UNIVERSIDAD CARLOS III MADRID

ESCUELA POLITÉCNICA SUPERIOR

DEPARTAMENTO DE TEORÍA DE LA SEÑAL Y COMUNICACIONES

GRADO EN INGENIERÍA DE SISTEMAS DE COMUNICACIONES

CURSO ACADÉMICO 2017 – 2018

TRABAJO FIN DE GRADO

DETECCIÓN ACÚSTICA DE EVENTOS VIOLENTOS

AUTORA: ELENA HIDALGO CUELLAS

TUTORA: CARMEN PELÁEZ MORENO

AGRADECIMIENTOS

En primer lugar, dar las gracias a mi tutora Carmen Peláez Moreno por permitirme realizar este Trabajo Fin de Grado, por toda su dedicación, ayuda y motivación semana tras semana.

A mis padres y a mi hermano Pablo por su apoyo incondicional a lo largo de estos duros años de carrera, manteniéndose a mi lado en todo momento.

A mis amigos y compañeros de carrera, Nuria, Alejandro, Patri, y especialmente a Antonio, por no dejarme sola en este camino y acompañarme en los momentos difíciles y sobre todo en los mejores momentos que me han dejado estos años.

RESUMEN

Los eventos violentos son situaciones que cada día escuchamos y vemos en la vida, sobre todo en películas a las que estamos acostumbrados a ver diariamente. Estas películas están clasificadas por edades en función a la cantidad de escenas de gran crueldad o violencia. Este es quizás la aplicación más común del tema que tratamos en este proyecto, pero no es el único. ¿Cómo sería poder predecir situaciones de violencia y poder avisar a los servicios de emergencia lo antes posible?

En este estudio buscamos el poder predecir ciertos eventos violentos en archivos de audio. Utilizando los datos proporcionados por la base de datos utilizada, extraeremos las características del audio a estudiar y tendremos que modelar los datos proporcionados para adecuarlos al estudio y así poder manejarlos de mejor manera. Para ello, nos basaremos en herramientas basadas en el aprendizaje máquina para intentar obtener unos buenos resultados de predicción. Dentro del aprendizaje máquina buscaremos los algoritmos o métodos que más se ajusten a las necesidades del proyecto, evaluando sobre todo aquellos relacionados con la clasificación.

ÍNDICE DE CONTENIDOS

1. INTRODUCCIÓN Y OBJETIVOS. 1

1.1. Motivación. 1

1.2. Objetivos. 3

1.3. Estructura del documento. 4

1.4. Marco regulador. 4

1.5. Impacto socio – económico. 5

2. ESTADO DEL ARTE Y METODOLOGÍA. 7

2.1. Estado del arte. 7

2.2. Metodología. 11

2.2.1. Aprendizaje Máquina. 12

3. IMPLEMENTACIÓN 18

3.1. Estudio bases de datos. 18

3.2. Herramientas utilizadas. 20

3.2.1. Matlab. 20

3.2.2. Lenguaje Python. 20

3.2.2.1. Spyder. 21

3.3. Modelado de los datos. 21

4. EXPERIMENTACIÓN 27

4.1. Película Soy Leyenda. 27

4.2. Película Salvar al Soldado Ryan 29

4.3. Película Piratas del Caribe y la Perla Negra. 31

4.4. Película Independence Day. 33

4.5. Película Fight Club. 35

4.6. Película Armageddon. 37

4.7. Película Eragon..... 39

4.8. Película Dead Poets Society. 40

4.9. Película Billy Elliot. 40

4.10. Película El Sexto Sentido. 41

4.11. Análisis de los resultados obtenidos. 42

5. CONCLUSIONES Y LÍNEAS FUTURAS..... 48

5.1. Conclusiones..... 48

5.2. Líneas futuras..... 49

6. BIBLIOGRAFÍA..... 50

7. ANEXO I – GRÁFICAS EXPERIMENTOS..... 54

7.1. Película Soy Leyenda. 54

7.2. Película Salvar al soldado Ryan..... 55

7.3. Película Piratas del Caribe y la Perla Negra. 57

7.4. Película Independence Day. 58

7.5. Película Fight Club..... 60

7.6. Película Armageddon. 61

7.7. Película Eragon..... 63

7.8. Película Dead Poets Society. 64

7.9. Película Billy Elliot. 64

7.10. Película El Sexto Sentido. 65

8. ANEXO II – PRESUPUESTO..... 66

9. ANEXO III - ABSTRACT. 68

9.1. Introduction and objetives. 68

9.2. Development of the proyect. 69

9.2.1. Theory. 69

9.2.1.1. Machine Learning. 69

9.2.2. Implementation..... 71

9.2.2.1. Databases..... 71

9.2.2.2. Modeling the data. 72

9.2.3. Experimentation and results. 72

9.3. Conclusion. 74

ÍNDICE DE TABLAS

Tabla 1 – Matriz de confusión del estudio "Audio surveillance using a bag of aural words classifier" 9

Tabla 2 – Conceptos Precision-Recall 16

Tabla 3 – Películas utilizadas en el proyecto 19

Tabla 4 – Película “Soy Leyenda” matriz de confusión etiqueta gritos 27

Tabla 5 – Película “Soy Leyenda” resultados etiqueta gritos 27

Tabla 6 – Película “Soy Leyenda” matriz de confusión etiqueta explosiones 28

Tabla 7 - Película “Soy Leyenda” resultados etiqueta explosiones 28

Tabla 8 – Película “Soy Leyenda” matriz de confusión etiqueta disparos 28

Tabla 9 - Película “Soy Leyenda” resultados etiqueta disparos 28

Tabla 10 – Película “Salvar al soldado Ryan” matriz de confusión etiqueta gritos 29

Tabla 11 - Película “Salvar al soldado Ryan” resultados etiqueta gritos 29

Tabla 12 - Película “Salvar al soldado Ryan” matriz de confusión etiqueta explosiones 30

Tabla 13 - Película “Salvar al soldado Ryan” resultados etiqueta explosiones 30

Tabla 14 - Película “Salvar al soldado Ryan” matriz de confusión etiqueta disparos 30

Tabla 15 - Película “Salvar al Soldado Ryan” resultados etiqueta disparos 31

Tabla 16 - Película “Piratas del Caribe y la perla negra” matriz de confusión etiqueta gritos 31

Tabla 17 - Película “Piratas del Caribe y la Perla Negra” resultados etiqueta gritos 32

Tabla 18 - Película “Piratas del Caribe y la perla negra” matriz de confusión etiqueta explosiones 32

Tabla 19 – Película “Piratas del Caribe y la Perla Negra” resultados etiqueta explosiones 32

Tabla 20 - Película “Piratas del Caribe y la perla negra” matriz de confusión etiqueta disparos 33

Tabla 21 – Película “Piratas del Caribe y la Perla Negra” resultados etiqueta disparos 33

Tabla 22 - Película “Independence Day” matriz de confusión etiqueta gritos 34

Tabla 23 – Película “Independence Day” resultados etiqueta gritos 34

Tabla 24 - Película “Independence Day” matriz de confusión etiqueta explosiones 34

Tabla 25 – Película “Independence Day” resultados etiqueta explosiones 34

Tabla 26 - Película “Independence Day” matriz de confusión etiqueta disparos 35

Tabla 27 – Película “Independence Day” resultados etiqueta disparos 35

Tabla 28 - Película “Fight Club” matriz de confusión etiqueta gritos 35

Tabla 29 – Película “Fight Club” resultados etiqueta gritos 35

Tabla 30 - Película “Fight Club” matriz de confusión etiqueta explosiones 36

Tabla 31 – Película “Fight Club” resultados etiqueta explosione 36

Tabla 32 - Película “Fight Club” matriz de confusión etiqueta disparos 36

Tabla 33 – Película “Fight Club” resultados etiqueta disparos 36

Tabla 34 - Película “Armageddon” matriz de confusión etiqueta gritos 37

Tabla 35 – Película “Armageddon” resultados etiqueta gritos 37

Tabla 36 - Película “Armageddon” matriz de confusión etiqueta explosiones 38

Tabla 37 – Película “Armageddon” resultados etiqueta explosiones 38

Tabla 38 - Película “Armageddon” matriz de confusión etiqueta disparos..... 38

Tabla 39 – Película “Armageddon” resultados etiqueta disparos..... 38

Tabla 40 - Película “Eragon” matriz de confusión etiqueta gritos 39

Tabla 41 – Película “Eragon” resultados etiqueta gritos 39

Tabla 42 - Película “Eragon” matriz de confusión etiqueta explosiones 39

Tabla 43 – Película “Eragon” resultados etiqueta explosiones 40

Tabla 44 - Película “Dead Poets Society” matriz de confusión etiqueta gritos..... 40

Tabla 45 – Película “Dead Poets Society” resultados etiqueta gritos..... 40

Tabla 46 - Película “Billy Elliot” matriz de confusión etiqueta gritos..... 41

Tabla 47 – Película “Billy Elliot” resultados etiqueta gritos..... 41

Tabla 48 - Película “El Sexto Sentido” matriz de confusión etiqueta gritos..... 41

Tabla 49 – Película “El Sexto Sentido” resultados etiqueta gritos..... 42

Tabla 50 - Película “El Sexto Sentido” matriz de confusión etiqueta disparos..... 42

Tabla 51 – Película “El Sexto Sentido” resultados etiqueta disparos 42

Tabla 52 – Resultados totales precisión, recall y *F-scores* para la etiqueta gritos. 43

Tabla 53 – Resultados totales etiqueta gritos, valores de la media, varianza y desviación típica. 44

Tabla 54 – Resultados totales precisión, recall y *F-scores* para la etiqueta explosiones. 44

Tabla 55 – Resultados totales etiqueta explosiones, valores de la media, varianza y desviación típica. 45

Tabla 56 – Resultados totales precisión, recall y *F-scores* para la etiqueta disparos..... 46

Tabla 57 – Resultados totales etiqueta disparos, valores de la media, varianza y desviación típica. 46

Tabla 58 – Costes de Personal 66

Tabla 59 – Costes de Material 66

Tabla 60 – Costes Generales 67

Tabla 61 – Presupuesto Total de Proyecto 67

Tabla 62 – Results Label Screams 73

Tabla 63 – Results Label Explosions 73

Tabla 64 – Results Label Gunshots 74

ÍNDICE DE FIGURAS

Figura 1 – Diagrama de bloques del estudio "Audio surveillance using a bag of aural words classifier" 10

Figura 2 – Diagrama de bloques proyecto. 11

Figura 3 – Cheat sheet Scikit-learn Algoritmos. Tomado de la referencia [18]..... 13

Figura 4 – Hiperplano SVM. Tomado de la referencia [2] 14

Figura 5 – Posibles Hiperplanos SVM. Tomado de la referencia [2]..... 14

Figura 6 – Hiperplano óptimo SVM. Tomado de la referencia [2] 15

Figura 7 – Película “Soy Leyenda” curvas *Precision-Recall* etiqueta gritos 54

Figura 8 – Película “Soy Leyenda” curvas *Precision-Recall* etiqueta explosiones 54

Figura 9 – Película “Soy Leyenda” curvas *Precision-Recall* etiqueta disparos 55

Figura 10 – Película “Salvar al soldado Ryan” curvas *Precision-Recall* etiqueta gritos 55

Figura 11 – Película “Salvar al soldado Ryan” curvas *Precision-Recall* etiqueta explosiones 56

Figura 12 – Película “Salvar al soldado Ryan” curvas *Precision-Recall* etiqueta disparos 56

Figura 13 – Película “Piratas del Caribe y la Perla Negra” curvas *Precision-Recall* etiqueta gritos 57

Figura 14 – Película “Piratas del Caribe y la Perla Negra” curvas *Precision-Recall* etiqueta explosiones..... 57

Figura 15 – Película “Piratas del Caribe y la Perla Negra” curvas *Precision-Recall* etiqueta disparos 58

Figura 16 – Película “Independence Day” curvas *Precision-Recall* etiqueta gritos 58

Figura 17 – Película “Independence Day” curvas *Precision-Recall* etiqueta explosiones 59

Figura 18 – Película “Independence Day” curvas *Precision-Recall* etiqueta disparos .. 59

Figura 19 – Película “Fight Club” curvas *Precision-Recall* etiqueta gritos 60

Figura 20 – Película “Fight Club” curvas *Precision-Recall* etiqueta explosiones 60

Figura 21 -– Película “Fight Club” curvas *Precision-Recall* etiqueta disparos 61

Figura 22 – Película “Armageddon” curvas *Precision-Recall* etiqueta gritos 61

Figura 23 – Película “Armageddon” curvas *Precision-Recall* etiqueta explosiones..... 62

Figura 24 – Película “Armageddon” curvas *Precision-Recall* etiqueta disparos 62

Figura 25 – Película “Eragon” curvas *Precision-Recall* etiqueta gritos..... 63

Figura 26 – Película “Eragon” curvas *Precision-Recall* etiqueta explosiones..... 63

Figura 27 – Película “Dead Poets Society” curvas *Precision-Recall* etiqueta gritos 64

Figura 28 – Película “Billy Elliot” curvas *Precision-Recall* etiqueta gritos 64

Figura 29 – Película “El Sexto Sentido” curvas *Precision-Recall* etiqueta gritos 65

Figura 30 – Película “El Sexto Sentido” curvas *Precision-Recall* etiqueta disparos 65

1. INTRODUCCIÓN Y OBJETIVOS.

En las siguientes líneas expondremos el problema que tratamos en este proyecto, así como las motivaciones que han ayudado a su desarrollo, los objetivos que se persiguen y por último hablaremos de su estructura.

Hoy en día las películas, las series, todo lo relacionado con el cine están al alcance de todos. Por ello es interesante saber y conocer los efectos o emociones que tendrán esas películas o series en nosotros. Ahora mismo todas las películas están calificadas en función de la edad, se supone que si una película es violenta o tiene escenas que pueden producir emociones fuertes en las personas tendrá una calificación edad alta. Pero el problema que vamos a resolver es: ¿cómo se califican estas películas? Es decir, cómo podemos detectar ciertas emociones en el audio de dichas películas. Pues bien, nos vamos a centrar en descubrir e intentar detectar los eventos violentos, los cuales son momentos en las películas que nos pueden producir miedo, ansiedad o angustia, por ejemplo: gritos, explosiones, disparos; en dichos archivos de audio.

1.1. Motivación.

La violencia es una de las causas o razones para la ordenación de las películas en función de la edad. La definición sobre este tema, que más se ajusta a este tema, es la que utiliza la Organización Mundial de la Salud (OMS); "violencia es el uso intencional de la fuerza o el poder físico, de hecho, o como amenaza, contra uno mismo, otra persona o un grupo o comunidad, que cause o tenga muchas probabilidades de causar lesiones, muerte, daños psicológicos, trastornos del desarrollo o privaciones". Esta definición del concepto habla de las consecuencias de la violencia, lo cual está muy relacionado con los eventos que queremos clasificar. Estos efectos pueden producir sentimientos o emociones en las personas, normalmente los eventos violentos pueden producir miedo o ansiedad, pero antes de introducirnos en ellos hablaremos sobre las emociones.

Una emoción está definida en la RAE como: "Alteración del ánimo intensa y pasajera, agradable o penosa, que va acompañada de cierta conmoción somática" o "Interés, generalmente expectante, con que se participa en algo que está

ocurriendo”. Las emociones pueden ser clasificadas de diferentes maneras, ya que hoy en día sigue el debate sobre cuál es la forma más precisa y clara de clasificarlas. Una de las formas más empleadas para su clasificación es la diferenciación entre emociones primarias o básicas y complejas o secundarias. Las emociones primarias o básicas son aquellas que son innatas en el ser humano y son producidas normalmente como respuesta a un estímulo como pueden ser por ejemplo los recuerdos, pensamientos, sentimientos, etcétera. Las emociones primarias se caracterizan por tener un carácter intenso y que no pueden ser controladas. Dentro de este grupo se encuentran la tristeza, la ira, la alegría, el miedo, la sorpresa o el asco. Una vez contempladas las emociones primarias, hablaremos de las secundarias. A estas emociones se les atribuye el nombre de complejas debido a que es más difícil desarrollarlas, no son naturales como las anteriores, necesitan un pequeño aprendizaje que se lleva a cabo a partir de las experiencias vividas por el ser humano. Dentro de este grupo, el cual es más numeroso que el anterior, podemos encontrar, por ejemplo, la culpa, los celos, el orgullo, la vergüenza y el amor.

Dentro de este proyecto, las emociones que más se ajustan al tema que tratamos son el miedo o la ansiedad. La emoción más reconocida de las comentadas anteriormente es el miedo, la cual es generada ante la percepción de una amenaza o peligro. Por ello, está muy ligada con el estímulo que la genera, asociamos ciertas cosas directamente con dicha emoción. Por otro lado, tenemos la ansiedad, la cual muchas veces es equivocada con el miedo. La diferencia entre estas dos emociones son las situaciones o estímulos que las producen, el miedo por su parte está producido por estímulos que consideramos peligrosos y que pueden hacernos temer por nuestra vida, mientras que la ansiedad es generada por situaciones que suponen una amenaza para nuestros intereses. Pero también hay que contemplar que esto depende de cada persona, no todas las personas sienten estos estímulos de la misma manera.

En línea con lo expuesto anteriormente la motivación para este trabajo ha sido poder investigar, a partir de archivos de audio, dichos eventos violentos. Conocer esas situaciones que nos producen emociones como el miedo o la angustia cuando vemos alguna película.

1.2. Objetivos.

El objetivo principal de este proyecto es la detección de eventos violentos en archivos de audio. Con esto queremos poder clasificarlos e intentar deducirlos a partir de unos registros de audio previamente clasificados, extrayendo sus características.

Los audios utilizados se corresponden con algunas películas conocidas. A partir de ellos se han extraído una serie de características comunes y principales del audio. A parte de estas características tenemos unos archivos en los cuales están clasificados los diferentes eventos que se pueden extraer de dichos audios.

Lo primero es ordenar y estudiar la base de datos obtenida, y a partir de ahí con la ayuda de Matlab convertir dichas características en variables más manejables u operativas. Una vez terminado este proceso, basándonos en técnicas de aprendizaje máquina, intentaremos detectar los eventos violentos acústicamente.

Por último, plantearemos y discutiremos los resultados obtenidos, estableceremos unas conclusiones y marcaremos unas líneas futuras a seguir.

Para el correcto desarrollo del proyecto y la obtención precisa de los resultados, se han seguido los siguientes puntos:

1. Analizar las diferentes bases de datos obtenidas, para el correcto estudio desarrollado en este proyecto.
2. Documentarse sobre las diferentes características del audio.
3. Realizar un estudio sobre aprendizaje máquina, profundizando en el tema de la clasificación de datos.
4. Elección de herramientas a utilizar en el desarrollo de los experimentos.
5. Realización de los experimentos.
6. Evaluación de los resultados obtenidos.

1.3. Estructura del documento.

En el capítulo 2 estudiaremos el estado del arte relacionado con el proyecto, viendo otros estudios, realizados con anterioridad, sobre el tema que tratamos. Aparte, explicaremos la metodología que hemos seguido en el desarrollo del proyecto.

En el capítulo 3 hablaremos sobre la implementación, expondremos las bases de datos y explicaremos el desarrollo del código utilizado en el proyecto.

En el capítulo 4 los resultados a los experimentos realizados mediante la matriz de confusión y los valores de precisión, *recall* y *F-scores*. Haremos un análisis de los resultados obtenidos.

En el capítulo 5 finalizaremos el proyecto con una conclusión y el estudio de líneas futuras.

1.4. Marco regulador.

La regulación que se aplica en este proyecto es la relacionada con el tratamiento de las bases de datos, así como la utilización de dichos datos.

Las bases de datos son un bien protegido por las leyes de propiedad intelectual, pero todavía se encuentra en pleno desarrollo. La propiedad intelectual se relaciona con las creaciones por parte de una persona o grupo de personas. Por ejemplo, invenciones, obras literarias y artísticas. La propiedad intelectual se divide en dos categorías: la propiedad industrial, la cual es la relacionada con las marcas, diseños industriales, patentes de invención; y los derechos de autor, que abarca las obras literarias, películas, música. En este proyecto utilizamos audios de películas, las cuales se ciernen por los derechos de autor, por ello vamos a profundizar un poco más sobre ellos.

Los derechos de autor, según el diccionario de la Real Academia Española de la Lengua, se definen como “derecho que la ley reconoce al autor de una obra intelectual o artística para autorizar su reproducción y participar en los beneficios que esta genere”. Existen diferentes clases de derechos de autor: derechos patrimoniales, derechos morales, derechos conexos, derechos de

reproducción, derecho de comunicación pública y los derechos de traducción. Dentro de las clases descritas, podríamos localizar los datos utilizados para el desarrollo del proyecto, dentro de la categoría de los derechos de reproducción, los cuales son aquellos que impiden a terceros efectuar copias o reproducciones de sus obras.

Las bases de datos pueden contener ciertos datos que conciernen a la reciente ley de protección de datos: Ley de Protección de Datos de Carácter Personal (GDPR). En esta ley es el reglamento relativo a la protección de personas físicas en lo que respecta al tratamiento de sus datos personales y la libre circulación de estos. Esta ley entró en vigor recientemente, en mayo de 2018. Dicha ley es una normativa a nivel europeo, por lo que concierne a todas las empresas, organismos o instituciones que manejen o posean datos personales de ciudadanos pertenecientes a cualquier país de la Unión Europea.

1.5. Impacto socio – económico.

Este proyecto se basa en los eventos violentos, los cuales hemos descrito anteriormente. Cuando hablamos de eventos violentos, estamos hablando de violencia. Este proyecto persigue detectar y ser capaz de predecir la violencia, con ello se podrían prevenir muchos momentos o situaciones violentas.

Este proyecto podría ser un inicio para el futuro desarrollo de un detector de violencia doméstica, donde se pueda prevenir algunas agresiones. Dicha patente podría ayudar a las personas que sufren malos tratos, por ejemplo, ya que se podrían instalar micrófonos en las casas de dichas personas y, en caso de detectar algún evento violento como un grito o un disparo, avisara mediante una alarma a los servicios de emergencia como pueden ser la policía y los servicios médicos. Así, se podría acudir al lugar con la mayor celeridad posible, ya que se ha detectado mediante un audio que se ha producido un evento violento que puede poner en peligro la vida de alguien.

Por otra parte, al igual que se podría instalar y desarrollar en casas, podría ser útil para detectar atentados en grandes lugares, y así avisar sin mediaciones a

los servicios de médicos y de emergencia para que acudan al lugar y así poder ayudar lo antes posible.

Por tanto, a partir de este proyecto se podría crear o desarrollar una patente que pudiera implantarse en los casos descritos anteriormente.

2. ESTADO DEL ARTE Y METODOLOGÍA.

2.1. Estado del arte.

En los estudios relacionados con el audio se necesita conocerlo previamente, estudiar sus características antes de experimentar con él. Por ello, en un primer lugar, vamos a exponer ciertas características del audio.

2.1.1. Métodos de extracción de características.

El *Spectral Centroid* (SC) es una medida utilizada para la caracterización del espectro durante el procesamiento de una señal digital. Indica donde se concentra el espectro, definiendo así su forma. Se utiliza para conocer la distribución de los componentes frecuenciales de una señal. Se calcula como la medida de las frecuencias de la señal, ponderada por las amplitudes, dividida por la suma de las amplitudes.

El flujo espectral (SF) o *Spectral Flux* es una medida que indica cuánto cambia la información espectral de una señal. Se obtiene calculando el cuadrado de la diferencia entre los espectros de dos tramas de audio consecutivas. Esta variación es un valor entre 0 y 1, si el valor obtenido es cercano al nulo, existe una gran similitud entre los espectros, y viceversa.

Zero Crossing Rate (ZCR) se define como la velocidad a la que la señal cambia de signo positivo a negativo. Esto indica que un valor de ZCR pequeño corresponde a una señal periódica, con poca presencia de ruido. El algoritmo para calcular el ZCR de una señal consiste en la comparación de dos muestras sucesivas, si son de signo contrario se computará como que la señal en ese intervalo tiene un cruce. Para finalizar ese valor se divide entre el número de muestras totales.

Los Coeficientes Cepstrales en las Frecuencias de Mel o MFCC (*Mel Frequency Cepstral Coefficients*) se utilizan para la representación del habla basándose en la percepción auditiva humana. Por ello son utilizados para el desarrollo de mecanismos de reconocimiento automático del habla. Esto es debido a que ayudan a la extracción de ciertas características de las componentes de la señal de audio, y dichas características sean adecuadas para la correcta identificación

de contenido relevante dentro de archivos de audio, así como obviar todas aquellas características que no tengan la suficiente importancia o que no aporten información valiosa para el reconocimiento. Dichos coeficientes fueron introducidos por *Davis* y *Mermelstein* en el año 1980 para ayudar al reconocimiento automático de la voz. Los coeficientes cepstrales son aquellos que derivan del cálculo del cepstrum de una señal. El cepstrum de una señal es lo que se conoce como la Transformada de Fourier del logaritmo del espectro de la señal previamente enventanada.

RMS (*Root Mean Square*) es la media cuadrática o valor cuadrático medio, es una medida estadística de la magnitud de la cantidad variable. Por ello es considerada como energía.

La tasa de error de bits o BER (*Bit Error Rate*) es la cantidad de bits que se reciben erróneamente durante una comunicación a través de un canal durante un periodo de tiempo dado. Estas alteraciones pueden producirse por el ruido, la interferencia o la distorsión que se producen en el canal.

El ancho de banda o BW nos indica la diferencia entre la frecuencia máxima y mínima en la que hay energía en la señal. El ancho de banda de una señal de audio está comprendido entre los 20 Hz y los 20 kHz. Este ancho de banda corresponde en grosso modo con el rango de frecuencias que pueden ser percibidas por el ser humano.

2.1.2. Estudios relacionados.

Una vez expuestas las principales características del audio, a continuación, veremos algunos estudios en los que los resultados se obtienen a partir del estudio y desarrollo de ciertos métodos que utilizan las características expuestas anteriormente.

Un estudio muy relacionado con este proyecto se titula: *"Audio surveillance using a bag of aural words classifier"* [1] y fue expuesto en 2013 en la *10th IEEE International Conference on Advanced Video and Signal Based Surveillance.* Este análisis se centra en la detección de eventos basados en audio,

para ello cuentan con diferentes sonidos de interés, los cuales están previamente clasificados, y un registro de audio. Estos registros de audio están compuestos por sonidos tomados por separado y combinados. El objetivo del sistema es encontrar si dentro de ese registro de audio se encuentra algún sonido de interés. Además de los sonidos clasificados, en dicho registro de audio se encuentran otros sonidos que consideran sonido de fondo. En el estudio se utilizan las características expuestas anteriormente, las cuales son cuantificadas con el algoritmo de *clustering* K-Means. Estas características son modeladas bajo el clasificador Support Vector Machine (SVM). Este estudio utiliza tres sonidos de interés: los gritos, los disparos y la rotura de cristales. Para la evaluación experimental han utilizado los siguientes valores: frecuencia de muestreo de 32 kHz, 1024 muestras por trama, 372 tramas, K=1024, donde K es el número de *clusters* o grupos, el cual es decidido como aquel que maximiza la exactitud final de la clasificación, 3 clases, las cuales son: disparos, gritos y rotura de cristales.

	BN	**BG**	**GS**	**S**
BN	0.961	0.032	0.004	0.003
BG	0.060	0.937	0.003	0.000
GS	0.021	0.011	0.968	0.000
S	0.030	0.005	0.000	0.965

Tabla 1 – Matriz de confusión del estudio "Audio surveillance using a bag of aural words classifier"

En la Tabla 1 podemos observar los pequeños errores de clasificación que se obtuvieron en dicho estudio, siendo BN (*Background Noise*) el sonido de fondo, BG (*Broken Glass*) corresponde al sonido producido por la rotura de los cristales, GS (*Gunshots*) son el sonido de los disparos y, por último, S (*Screams*) son los gritos. Dichos errores son alrededor de un 3%, lo cual se puede considerar prácticamente despreciable. Con estos valores y utilizando el algoritmo SVM se obtiene una exactitud del 95.8%, afirmando así la validez del estudio.

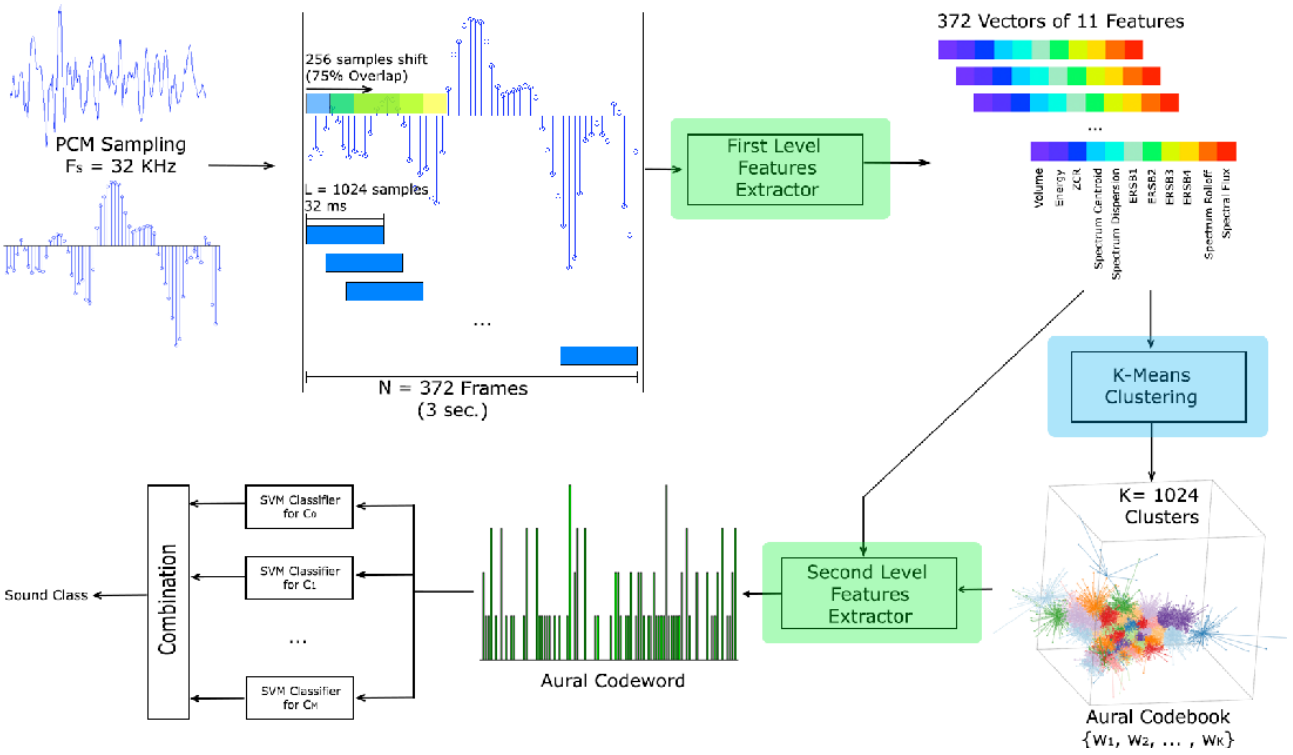

Figura 1 – Diagrama de bloques del estudio "Audio surveillance using a bag of aural words classifier"

En la Figura 1 se expone el diagrama de bloques de dicho estudio, en el que se aprecian cada una de las fases que se realizaron para obtener los resultados.

En segundo lugar, otro estudio a tener en cuenta es el llevado a cabo por *MediaEval Benchmarking Initiative for Multimedia Evaluation*. MediaEval es una iniciativa dedicada a la evaluación de nuevos algoritmos para el acceso y recuperación de archivos multimedia, destaca su enfoque sobre el análisis del audio y el reconocimiento vocal. Para realizar estos estudios utilizan la misma base de datos que hemos utilizado para el desarrollo de este proyecto y que describimos en el apartado 3.1, la base de datos *Technicolor*. Para el desarrollo del estudio llevado a cabo por MediaEval, transforman el estudio en un concurso, donde los participantes se organizan en grupos e intentan desarrollar el mejor algoritmo posible para la clasificación de eventos o análisis del audio. Es decir, MediaEval proporciona la base de datos sobre la que deben trabajar los diferentes grupos inscritos. Además, se proponen varios temas donde los participantes pueden desarrollar sus algoritmos. Una vez que se han obtenido dichos algoritmos se organiza una convención donde se exponen, sintetizan y

evalúan. Dependiendo del año en el que se realice la convención o concurso varía el enfoque sobre el que se desarrollan los algoritmos. Por ejemplo, en el año 2016, entre los temas tratados y propuestos, se pedía el desarrollo de predicciones relacionadas con el impacto que ocasionan ciertos archivos multimedia en las personas. Se requería el desarrollo de un algoritmo que predijera las emociones suscitadas o provocadas en los espectadores.

Una vez expuestos estos estudios, para poder desarrollar este proyecto hemos seguido ciertos pasos, los cuales representamos en la *Figura 2* en un diagrama de bloques.

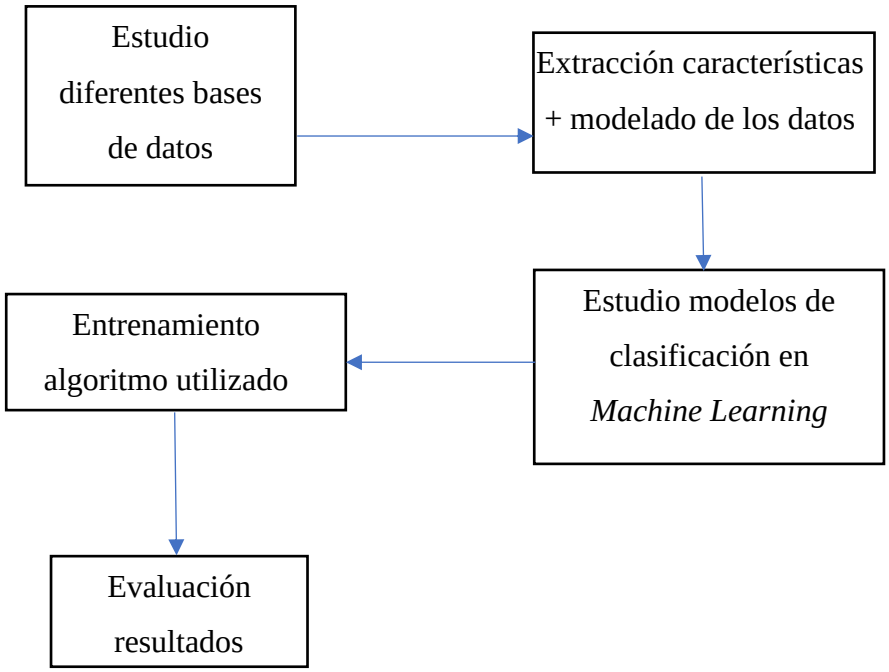

Figura 2 – Diagrama de bloques proyecto.

2.2. Metodología.

En este apartado detallaremos los mecanismos implementados para el desarrollo del proyecto, explicaremos la elección de cada de ellos, así como la elección de los métodos de clasificación de aprendizaje máquina para el modelado de los datos.

2.2.1. Aprendizaje Máquina.

El aprendizaje automático o máquina es un tipo de inteligencia artificial que desarrolla en las máquinas u ordenadores la capacidad de aprender, sin ser programadas explícitamente. Se centra en el desarrollo de algoritmos que otorguen a las máquinas la capacidad de poder encontrar ciertos patrones o comportamientos, a partir de ciertos datos o ejemplos dados previamente. Una posible clasificación de estos algoritmos sería: supervisados o no supervisados.

Los algoritmos supervisados o aprendizaje supervisado son aquellos que, en base a lo aprendido sobre ciertos datos etiquetados previamente, son capaces de etiquetar correctamente un dato a la salida del sistema, es decir, es capaz de predecir la etiqueta de salida. Este aprendizaje suele utilizarse en problemas de clasificación y regresión. Dentro de este grupo se encuentran algoritmos como los árboles de decisión, la regresión por mínimos cuadrados, máquinas de vectores soporte (SVM) o la regresión logística.

Los algoritmos no supervisados o aprendizaje no supervisado tienen lugar cuando no se posee de datos etiquetados previamente, no existe un conocimiento a priori de dichas etiquetas. Se conocen los datos de entrada al sistema, pero no existen datos de salida relacionados con los de entrada como ocurría en el aprendizaje supervisado. Para este desarrollo se tiene que encontrar algún tipo de organización o clasificación que ayude a simplificar el análisis. Se debe encontrar una estructura de datos mediante observaciones. Estos algoritmos se utilizan en problemas de *clustering* o agrupación principalmente. A este grupo pertenecen algoritmos como los algoritmos de descomposición en valores singulares (*Singular Value Decomposition*, SVD), el análisis de componentes independientes (*Independent Component Analysis*, ICA) y, los ya mencionados, algoritmos de clustering.

La elección del algoritmo a utilizar es compleja y depende de los datos con lo que iniciemos el estudio. Para agilizar y ayudar en dicha elección existen las llamadas *cheat-sheet*. Un ejemplo de ellas es la siguiente ilustración:

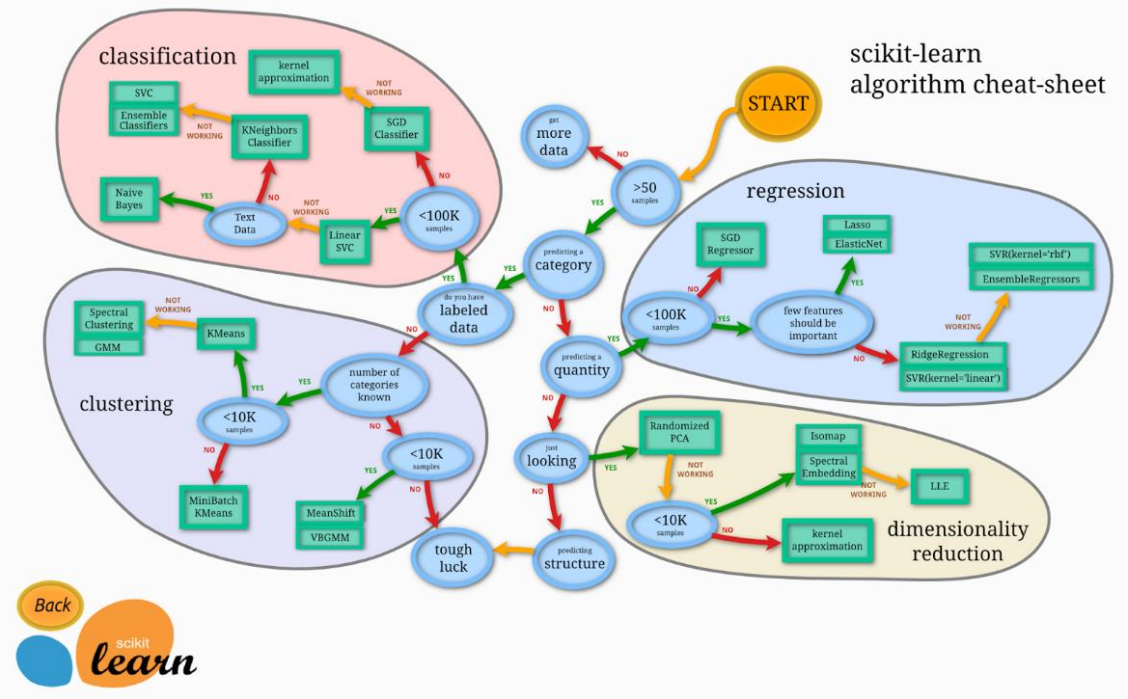

Figura 3 – Cheat sheet Scikit-learn Algoritmos. Tomado de la referencia [18]

Siguiendo los pasos que exponen en la Figura 3, decidimos utilizar el algoritmo SVM debido a que posee funciones que se adaptan y ayudan a la obtención de los resultados del proyecto.

SVM es un conjunto de algoritmos de aprendizaje supervisado relacionados con problemas de clasificación y regresión. Dado un conjunto de muestras de entrenamiento se etiquetan las clases que existen en ella y a partir de ellas se entrena una SVM que prediga la clase de una nueva muestra. La SVM busca un hiperplano que separe de forma óptima a los puntos de una clase u otra. Aquí, en el concepto de "separación óptima", es donde reside la característica fundamental de las SVM, el hiperplano buscado será aquel que tenga una separación cuya distancia sea máxima respecto a los puntos que estén más cerca de él mismo. La siguiente figura muestra este concepto de hiperplano.

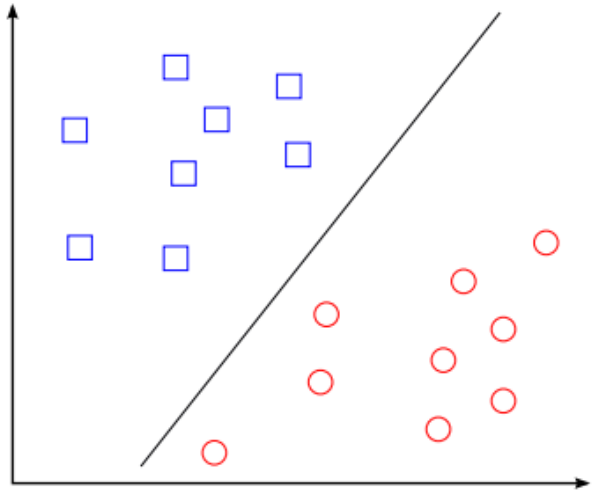

Figura 4 – Hiperplano SVM. Tomado de la referencia [2]

Pero como se puede observar este hiperplano no es único como demuestra la siguiente figura.

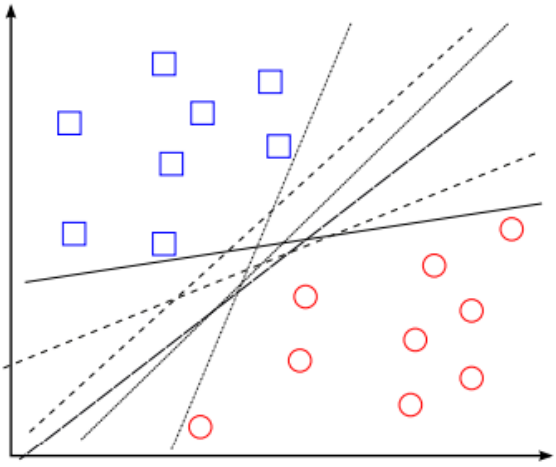

Figura 5 – Posibles Hiperplanos SVM. Tomado de la referencia [2]

A partir de este concepto, que el hiperplano no es único, surge la pregunta sobre si es posible obtener un hiperplano de separación óptima. Para ello se debe definir el concepto de margen de un hiperplano, como la mínima distancia entre dicho hiperplano y la muestra más cercana de cualquiera de las dos

clases. Por tanto, el hiperplano será óptimo cuando el margen toma su tamaño máximo.

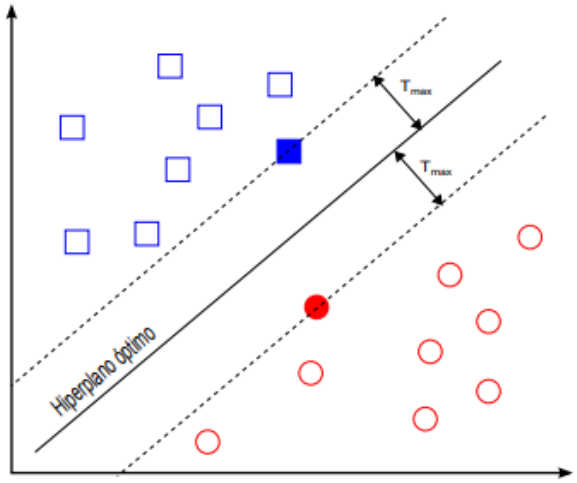

Figura 6 – Hiperplano óptimo SVM. Tomado de la referencia [2]

Una vez definido el algoritmo SVM, utilizamos las curvas *precision – recall* para visualizar la clasificación de dicho algoritmo. Estas curvas se utilizan para evaluar la calidad de salida del clasificador. En la recuperación de información, la precisión es una medida de relevancia del resultado, mientras que *recall* es una medida de cuántos resultados realmente se devuelven. Para entenderlo mejor, tenemos que definir previamente algunos conceptos:

- TN (True Negatives) o Verdaderos Negativos: El número de casos en los que fue predicho negativo y era negativo.
- TP (True Positives) o Verdaderos Positivos: El número de casos en los que fue predicho positivo y era positivo.
- FN (False Negatives) o Falsos Negativos: El número de casos en los que fue predicho negativo y era positivo.
- FP (False Positives) o Falsos Positivos: El número de casos en los que fue predicho positivo y era negativo.

En la siguiente tabla podemos visualizar estos conceptos.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Tabla 2 – Conceptos Precision-Recall

Por tanto, una vez vistos estas definiciones, vamos a aclarar los conceptos de *precisión* y *recall*.

La *precisión* para una clase es la proporción de elementos etiquetados como pertenecientes a la clase positiva, y que realmente pertenecen a ella. Esto se define en la siguiente fórmula:

$$Precisión = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall es la proporción de positivos reales que se identificaron correctamente. Esto se resume en la siguiente ecuación:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Otro parámetro que se utiliza para evaluar la precisión de un modelo de clasificación es *F-scores*. Se define como el promedio o media de precisión y *recall*. *F-scores* toma valores entre cero y uno, siendo un valor cercano o igual a uno el correspondiente a una clasificación correcta, donde la precisión y *recall* es casi perfecta. Si toma un valor cercano o igual a cero, el modelo no es correcto. Se define con la siguiente fórmula:

$$F\ scores = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

La elección de este algoritmo por delante de otros métodos que pertenecen al grupo de aprendizaje supervisado, como son los algoritmos de regresión por mínimos cuadrados, árboles de decisión o regresión logística, se debe a que en primer lugar buscamos un algoritmo de clasificación, no de regresión, por tanto, podemos reducir el número de algoritmos posibles. Dentro de los métodos de clasificación, el SVM posee muchas funciones que le diferencian de los demás, las cuales son muy importantes para el desarrollo de este proyecto, como por ejemplo la posibilidad de utilizar diferentes *kernel*, además permite el manejo de grandes cantidades de datos.

3. IMPLEMENTACIÓN

En este apartado detallaremos el proceso de obtención de los resultados del proyecto, hemos expuesto en el diagrama de bloques en la Figura 2.

3.1. Estudio bases de datos.

Al inicio del proyecto se obtuvieron diferentes bases de datos que serían de ayuda para la extracción de ciertos datos que permitieran el estudio que hemos desarrollado. En total se manejaron cinco bases de datos, relacionadas con el estudio de eventos violentos.

- Base de datos TECHNICOLOR [10][11][12][13]: compuesta por 25 películas y 86 vídeos de YouTube. Incluye un fichero que contiene cuándo se producen eventos violentos en la película, así como dos ficheros con características propias de cada película.
- Base de datos MAHNOB-HCI [21]: compuesta por 20 extractos de películas, cuya duración oscila entre los 35 y los 118 segundos. Contiene etiquetas relacionadas con las emociones, excitación y previsibilidad combinadas con vídeos de las expresiones faciales de personas. Estos datos fueron recogidos en un total de 30 participantes mientras visualizaban las películas y vídeos proporcionados por la base de datos.
- Base de datos DEAP [20]: contiene 120 vídeos de aproximadamente un minuto de duración. Es una base de datos multimodal para el análisis de los estados de ánimo de los seres humanos. A través de un electroencefalograma y las señales fisiológicas de 32 participantes, caracterizan las emociones producidas en las personas y su nivel.
- Base de datos HUMAINE [19]: compuesta por 50 vídeos con una duración que oscila entre los 5 segundos y los 3 minutos. En ella se etiquetan estados emocionales, eventos clave, palabras relacionadas con las emociones.
- Base de datos LIRIS-ACCEDE [22]: compuesta por 9800 extractos de 160 películas, con una duración entre los ocho y los doce segundos.

Una vez estudiadas cada una de las bases de datos, elegimos una de ellas. Decidimos realizar el proyecto utilizando la base de datos TECHNICOLOR, ya que es la que más se ajusta a los datos que necesitamos, ya que contiene varios ficheros que nos ayudan a desarrollar el proyecto. De esta base de datos obtuvimos varios ficheros, por cada película, en los que se etiquetaban los diferentes eventos que ocurrían en la película, en este caso los eventos etiquetados eran: sangre, persecuciones de coches, explosiones, peleas, fuego, armas de fuego, situaciones sangrientas, disparos y gritos. Además de estos interesantes ficheros, la base de datos también contenía un fichero donde se encontraban características del audio de cada película, estas características son el RMS, MFCC, BER, ZCR entre otras. Por ello, decidimos utilizar esta base de datos, así nos ayudaría a obtener los resultados que buscábamos.

Dentro de esta base de datos se encontraban más de veinte películas, pero para realizar el estudio seleccionamos las siguientes:

Película	**Duración**
Independence Day	2 horas y 33 minutos
Soy leyenda	1 hora y 44 minutos
Fight Club	2 horas y 31 minutos
Eragon	1 hora y 44 minutos
Dead Poets Society	2 horas y 20 minutos
Billy Elliot	1 hora y 51 minutos
Armageddon	2 horas y 33 minutos
Piratas del Caribe y la Perla Negra	2 horas y 23 minutos
El sexto sentido	1 hora y 50 minutos
Salvar al Soldado Ryan	2 horas y 50 minutos

Tabla 3 – Películas utilizadas en el proyecto

3.2. Herramientas utilizadas.

A continuación, expondremos las herramientas utilizadas en el desarrollo del proyecto, explicando las razones de su elección.

3.2.1. Matlab.

Matlab ("MAtrix LABoratory") es un entorno de desarrollo integrado con un lenguaje de programación propio (.m). Este software matemático permite realizar cálculos numéricos, así como, la visualización de estos. Dentro de Matlab, destacan sus integraciones para realizar análisis numérico, cálculo matricial, procesamiento de señales y visualización gráfica. Matlab es utilizado como herramienta de enseñanza, pero también se utiliza en áreas como la industria, para la investigación y la resolución de problemas prácticos. Matlab proporciona una serie de utilidades específicas denominadas Toolboxes, que existen para cada áerea del campo de la ingeniería y de la simulación. Con las Toolboxes son funciones que resuelven problemas como, por ejemplo, el procesamiento de señales, identificación de sistemas, diseño de sistemas de control, redes neuronales.

Matlab ha sido elegida para el desarrollo del proyecto por las herramientas y funciones que ofrece en el tratamiento del audio, así como la gran capacidad para el tratamiento de grandes vectores de datos. Otro aspecto que se tuvo en cuenta para su elección fue la representación gráfica que permite dicho software.

3.2.2. Lenguaje Python.

Python es un lenguaje de programación utilizado para el desarrollo de aplicaciones de aprendizaje automático. Este lenguaje fue creado a finales de los años ochenta por Guido van Rossum en el CWI (Centrum Wiskunde & Informatica) en los Países Bajos. Una característica importante de este lenguaje es su sencilla sintaxis, además no requiere compilación, por lo que es utilizado en el desarrollo de proyectos de investigación y experimentación, debido a la facilidad para introducir cambios en el código sin invertir demasiado tiempo en las compilaciones. Su principal punto

fuerte es la introducción de librerías, como por ejemplo las librerías Numpy, las cuales ayudan a la generación de modelos de aprendizaje automático y redes neuronales. Otra característica importante que ofrece el lenguaje Python es su capacidad para manejar grandes ficheros de datos, realizando tareas en tiempos más que aceptables.

Las características expuestas son las que han producido la utilización de este lenguaje para el desarrollo del proyecto.

3.2.2.1. Spyder.

Spyder es un entorno de desarrollo interactivo para el lenguaje Python, posee una combinación de funcionalidades de edición, depuración, análisis, además de exploración de datos, ejecución y capacidades de visualización gráfica. Ofrece la integración de muchas librerías utilizadas al programar en Python, como son NumPy, SciPy, Matplotlib, entre otras.

La elección de esta herramienta para el desarrollo del proyecto por delante de otras aplicaciones como Jupyter Notebook, se debe a que este entorno de desarrollo ofrece una mayor facilidad para depurar el código y poder encontrar fallos en el mismo.

3.3. Modelado de los datos.

Una vez decidida la base de datos que vamos a utilizar, como hemos comentado en el apartado anterior, utilizaremos la base de datos TECHNICOLOR.

En dicha base de datos tenemos varios ficheros *.txt* en los cuales aparecen etiquetadas ciertos eventos que se han producido a lo largo de la película. Tenemos un total de once ficheros, pero no todos nos interesan. Por ello, decidimos hacer otra pequeña selección y quedarnos con los ficheros que más se ajusten a los datos que necesitamos. Por tanto, utilizamos los ficheros que tienen un reflejo acústico, los cuales son los gritos, los disparos y las explosiones.

En estos ficheros seleccionados, los eventos, vienen etiquetados por segundos, es decir, cada fichero contiene en qué segundos de la película seleccionada aparece el evento que estamos caracterizando.

En primer lugar, como los datos de los ficheros *.txt* se encuentran en segundos, para un mejor desarrollo del proyecto decidimos pasar los segundos a tramas, que almacenaremos en vectores. Para ello decidimos utilizar 40 ms/trama. Antes de realizar estos cálculos, hemos considerado que cuando existe un evento violento, en el vector, le pondremos un 1, y un 0 cuando no haya evento violento.

Para obtener muestras tenemos que multiplicar el inicio y el final del fichero en segundos por 44100, es decir:

$$Inicio\ fichero\ en\ segundos\ x\ 44100\frac{muestras}{segundo} = muestra\ inicial$$

$$Final\ fichero\ en\ segundos\ x\ 44100\frac{muestras}{segundo} = muestra\ final$$

Una vez que tenemos estas muestras, las agrupamos en grupos de 1764 muestras, el tamaño de trama es 40 ms y para tener 44100 muestras/segundo necesitamos grupos de 1764 muestras, obteniendo finalmente el número de tramas que buscábamos:

$$\frac{Muestras}{1764\frac{muestras}{trama}} = número\ de\ tramas$$

Una vez realizados estos cálculos, los cuales fueron realizados con la herramienta descrita anteriormente MATLAB, ya tenemos los vectores de

tramas sobre los que vamos a trabajar en el proyecto. A continuación, se muestra el código donde se realizan dichos cálculos.

```
inicio_muestras(k) = C1(k) * 44100;
inicio_muestras_agrupadas(k) = round(inicio_muestras(k)/1764);

final_muestras(j) = C2(j) * 44100;
final_muestras_agrupadas(j) = round(final_muestras(j)/1764)-1;
```

Además de estos vectores, tenemos una matriz con las características del audio que estamos estudiando. Dichas características son AE, RMS, BER, ZCR, SF, SC, BW, MFCC; que hemos descrito con anterioridad en el 2.1. Para el modelaje del algoritmo, utilizamos como variable X estas características acústicas, las cuales son proporcionadas por la base de datos.

A continuación, una vez tenemos los vectores de etiquetas de cada evento violento, gritos, explosiones y disparos, en tramas en la herramienta MATLAB, utilizamos la función *jsonencode* para exportar cada uno de los ficheros y utilizarlos en el software *Spyder*. Además de convertir los vectores de etiquetas, también utilizamos la función *csvwrite* para exportar el vector que contiene las características.

En tercer lugar, nos movemos a la herramienta *Spyder* para continuar con la programación en Python del algoritmo SVM. Los datos que hemos modelado están desequilibrados, es decir, hay muchas más muestras de una clase que de otra. Por ello, necesitamos equilibrar dichas muestras entre ambas clases, porque si no los resultados obtenidos son incorrectos. Para solucionar este problema acudimos a ciertas funciones para corregir el desequilibrio de los datos, para ello hay dos técnicas principales: *Over Sampling* y *Under Sampling*. *Over Sampling* consiste en aumentar el número de muestras de la clase con menor número de muestras hasta conseguir que ambas clases tengan el mismo número. *Under Sampling*, sin embargo, se basa en reducir el número

de muestras de la clase con mayor cantidad, hasta que tiene el mismo número de muestras que la otra clase.

La técnica que hemos utilizado en este proyecto es la de *Under Sampling*, para ello hemos utilizado la función *RandomUnderSampler*. En esta función equilibra las muestras de las dos clases, reduciendo los datos de la clase con mayor número de muestras. Una vez que reducimos ambas clases, continuamos con el entrenamiento del algoritmo seleccionado.

Para el entrenamiento del algoritmo SVM, dividimos los datos que vamos a utilizar en conjuntos de *train* y *test*. El conjunto de *train* representará el 80% y el de *test* el 20%.

Utilizamos el siguiente código, para cada tipo de evento violento, para el entrenamiento del algoritmo:

```
data = pd.read_csv('características.csv')
etiquetas = json.loads(open('etiquetas.json').read())

X = np.array(data)
Y = np.array(etiquetas)
if (len(Y) < len(X)):
    diferencia = len(X) - len(Y)
    array_diferencia = np.zeros(diferencia, dtype=int)
    Y = np.concatenate((Y, array_diferencia))

rus = RandomUnderSampler(random_state=0)
rus.fit(X, Y)
X_resampled, Y_resampled = rus.sample(X, Y)

X_train, X_test, Y_train, Y_test =
train_test_split(X_resampled, Y_resampled, test_size = 0.20)
```

En este algoritmo utilizamos un *kernel* lineal, para ello utilizamos la función *LinearSVC, Linear Support Vector Classification*, la cual pertenece a la librería de funciones del algoritmo SVM en *sklearn*. A continuación, se muestran las líneas del código donde se entrena al algoritmo:

```
classifier = svm.LinearSVC(class_weight='balanced')
classifier.fit(X_train, Y_train)
y_pred = classifier.predict(X_test)

print(confusion_matrix(Y_test,y_pred))
print(classification_report(Y_test,y_pred))
```

En este código, en primer lugar, se entrena el modelo que utilizamos y luego realizamos predicciones. Para observar los resultados utilizamos la matriz de confusión y los valores de precisión, *recall* y *F-scores*. La matriz de confusión nos muestra los valores de los conceptos expuestos en el apartado 2.2.1: *True Positives, True Negatives, False Positives, False Negatives*. También en dicho apartado se explican los conceptos de precisión, *recall* y *F-scores*.

A continuación, para entender y observar los resultados más visualmente, utilizamos las curvas *precisión-recall*. En estas curvas se representan los valores de precisión y *recall* obtenidos, se utilizan para ponderar el acierto del algoritmo entrenado. Para ello utilizamos el siguiente código:

```
y_scores = classifier.decision_function(X_test)
precision_kernel,recall_kernel,_=
precision_recall_curve(Y_test, y_scores)
plt.step(recall_kernel, precision_kernel, color='b',
alpha=0.2,where='post')
plt.fill_between(recall_kernel, precision_kernel, step='post',
alpha=0.2,color='b')
```

```
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.title("SCREAMS")
plt.show()
```

Por último, calculamos el área bajo la curva obtenida, para tener una visión más clara de los resultados obtenidos. Utilizamos la función *auc.*

```
auc = metrics.auc(recall_kernel, precision_kernel)
print(auc)
```

Una vez explicado el código, vamos a explicar los resultados obtenidos para cada película.

4. EXPERIMENTACIÓN

En este apartado expondremos los resultados obtenidos al realizar los experimentos, para ello nos apoyaremos en los conceptos *precisión*, *recall* y *F-scores*, explicados previamente en el apartado 2.2.1. En este estudio consideramos dos clases: 1, si existe evento violento, y 0, si no existe.

En el ANEXO I se encuentran las gráficas de las curvas *precisión-recall* y los valores del área bajo dichas curvas correspondientes a estos experimentos.

4.1. Película Soy Leyenda.

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	2206	91
	1	1659	664

Tabla 4 – Película "Soy Leyenda" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.73	0.62	0.57	4620

Tabla 5 – Película "Soy Leyenda" resultados etiqueta gritos

- Explosiones:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	130	11
	1	22	115

Tabla 6 – Película "Soy Leyenda" matriz de confusión etiqueta explosiones

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.88	0.88	0.88	278

Tabla 7 - Película "Soy Leyenda" resultados etiqueta explosiones

- Disparos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	164	26
	1	59	162

Tabla 8 – Película "Soy Leyenda" matriz de confusión etiqueta disparos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.80	0.79	0.79	411

Tabla 9 - Película "Soy Leyenda" resultados etiqueta disparos

Observando los resultados obtenidos para los tres tipos de eventos violentos estudiados, vemos que se obtienen unos mejores resultados para la etiqueta explosiones. Pero aun así los resultados de las otras etiquetas son aceptables.

4.2. Película Salvar al Soldado Ryan

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	2343	2085
	1	2395	2006

Tabla 10 – Película "Salvar al soldado Ryan" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.49	0.49	0.49	8829

Tabla 11 - Película "Salvar al soldado Ryan" resultados etiqueta gritos

Para esta etiqueta obtenemos unos valores bastante malos en la *precisión*, el *recall* y *F-scores*. Esto puede ser una consecuencia de una mala detección del algoritmo, debida a una mala agrupación de los eventos en los vectores de etiquetas. Ciertos eventos pueden durar más de lo que realmente duran en la película porque no lo hemos reagrupado correctamente.

- Explosiones:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	2278	3885
	1	1033	5106

Tabla 12 - Película "Salvar al soldado Ryan" matriz de confusión etiqueta explosiones

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.63	0.60	0.58	12302

Tabla 13 - Película "Salvar al soldado Ryan" resultados etiqueta explosiones

En la etiqueta de las explosiones no se obtienen unos resultados tan bajos como en las otras etiquetas de esta película, pero aún así no son óptimos. Esto puede darse por lo comentado anteriormente, mala reagrupación.

- Disparos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	1613	10991
	1	491	11992

Tabla 14 - Película "Salvar al soldado Ryan" matriz de confusión etiqueta disparos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.64	0.54	0.45	25087

Tabla 15 - Película "Salvar al Soldado Ryan" resultados etiqueta disparos

En este caso, la etiqueta disparos, obtiene unos valores para el *recall y F-scores* que no son muy buenos. En la matriz de confusión vemos que el algoritmo no clasifica correctamente muchos de los valores que deberían llevar la etiqueta 0. Esto puede deberse a lo comentado anteriormente, que los vectores de etiquetas no están agrupados correctamente. También se observa que son muchos datos los que se manejan, un total de 25087 valores, por tanto, al tener tantas muestras no están bien definidos las clases en los vectores de tramas que tratamos.

4.3. Película Piratas del Caribe y la Perla Negra.

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	3970	123
	1	3437	823

Tabla 16 - Película "Piratas del Caribe y la perla negra" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.71	0.57	0.50	8353

Tabla 17 - Película "Piratas del Caribe y la Perla Negra" resultados etiqueta gritos

Los valores obtenidos de *recall y F-scores* no son óptimos, el algoritmo no está clasificando correctamente la clase 1, es decir, cuando hay un evento. Esto podría darse al haber reducido el número de muestras que tratamos, al utilizar la función de *Under Sampling,* igualamos el número de muestras de la clase 0 a la de la clase 1.

- Explosiones:

Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	266	57
	1	40	273

Tabla 18 - Película "Piratas del Caribe y la perla negra" matriz de confusión etiqueta explosiones

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.85	0.85	0.85	636

Tabla 19 – Película "Piratas del Caribe y la Perla Negra" resultados etiqueta explosiones

En la predicción de etiquetas de explosiones, observamos en la Tabla 19, que los valores son buenos, el algoritmo clasifica correctamente el 85% de los casos.

- Disparos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	783	28
	1	712	116

Tabla 20 - Película "Piratas del Caribe y la perla negra" matriz de confusión etiqueta disparos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.67	0.55	0.46	1639

Tabla 21 – Película "Piratas del Caribe y la Perla Negra" resultados etiqueta disparos

Los valores obtenidos de *recall* y *F-scores* no son valores buenos, el clasificador no funciona correctamente. Esto puede deberse a que teníamos muy pocos valores de la clase 1, y al equilibrar los datos, no se corresponde correctamente el número de etiquetas que teníamos con el que tenemos ahora.

4.4. Película Independence Day.

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	1596	269
	1	1243	620

Tabla 22 - Película "Independence Day" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.63	0.59	0.56	3728

Tabla 23 – Película "Independence Day" resultados etiqueta gritos

En este caso pasa lo mismo que hemos expuesto en el 4.3 en el apartado de disparos. Es la clase 1 la que no se clasifica correctamente.

- Explosiones:

Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	935	169
	1	117	1003

Tabla 24 - Película "Independence Day" matriz de confusión etiqueta explosiones

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.87	0.87	0.87	2224

Tabla 25 – Película "Independence Day" resultados etiqueta explosiones

- Disparos:

Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	469	140
	1	20	599

Tabla 26 - Película "Independence Day" matriz de confusión etiqueta disparos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.88	0.87	0.87	1228

Tabla 27 – Película "Independence Day" resultados etiqueta disparos

Para las etiquetas de las explosiones y los disparos los resultados obtenidos son relativamente buenos, el algoritmo clasifica válidamente.

4.5. Película Fight Club.

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	1882	112
	1	1521	393

Tabla 28 - Película "Fight Club" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.66	0.58	0.51	3908

Tabla 29 – Película "Fight Club" resultados etiqueta gritos

Este es otro caso en el que se clasifica mal la clase 1, la explicación es lo expuesto en el apartado 4.3 en la sección de disparos.

- Explosiones:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	55	38
	1	1	99

Tabla 30 - Película "Fight Club" matriz de confusión etiqueta explosiones

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.85	0.80	0.79	193

Tabla 31 – Película "Fight Club" resultados etiqueta explosione

- Disparos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	27	3
	1	16	12

Tabla 32 - Película "Fight Club" matriz de confusión etiqueta disparos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.71	0.67	0.65	58

Tabla 33 – Película "Fight Club" resultados etiqueta disparos

Estos resultados pueden darse debido a lo expuesto en los apartados anteriores, concretamente en el apartado 4.3 en la sección de disparos.

4.6. Película Armageddon.

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	618	1263
	1	108	1736

Tabla 34 - Película "Armageddon" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.72	0.63	0.59	3725

Tabla 35 – Película "Armageddon" resultados etiqueta gritos

En este caso se predice mal la clase 0, es decir, cuando no hay evento violento. Esto puede darse debido a una mala reagrupación, como hemos comentado en los apartados anteriores, como por ejemplo en el apartado 4.2 en la sección de gritos.

- Explosiones:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	1146	1348
	1	17	2463

Tabla 36 - Película "Armageddon" matriz de confusión etiqueta explosiones

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.82	0.73	0.70	4974

Tabla 37 – Película "Armageddon" resultados etiqueta explosiones

- Disparos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	125	50
	1	5	182

Tabla 38 - Película "Armageddon" matriz de confusión etiqueta disparos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.87	0.85	0.85	362

Tabla 39 – Película "Armageddon" resultados etiqueta disparos

4.7. Película Eragon.

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	163	1662
	1	45	1805

Tabla 40 - Película "Eragon" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.65	0.54	0.42	3675

Tabla 41 – Película "Eragon" resultados etiqueta gritos

- Explosiones:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	76	58
	1	36	86

Tabla 42 - Película "Eragon" matriz de confusión etiqueta explosiones

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.64	0.63	0.63	256

Tabla 43 – Película "Eragon" resultados etiqueta explosiones

En esta clasificación de las etiquetas de la película Eragon, observamos que ocurre lo mismo que hemos expuesto anteriormente el algoritmo no clasifica correctamente cierta clase.

4.8. Película Dead Poets Society.

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	1223	7
	1	1090	103

Tabla 44 - Película "Dead Poets Society" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.73	0.55	0.43	2423

Tabla 45 – Película "Dead Poets Society" resultados etiqueta gritos

En este caso, observamos que vuelve a clasificar mal la clase 1, por lo ya expuesto anteriormente.

4.9. Película Billy Elliot.

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	1038	485
	1	311	1262

Tabla 46 - Película "Billy Elliot" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.75	0.74	0.74	3096

Tabla 47 – Película "Billy Elliot" resultados etiqueta gritos

Para esta película obtenemos unos valores bastante buenos, por lo que el clasificador realiza su función.

4.10. Película El Sexto Sentido.

Estos son los resultados obtenidos para cada uno de los eventos violentos clasificados:

- Gritos:

  Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	285	125
	1	102	290

Tabla 48 - Película "El Sexto Sentido" matriz de confusión etiqueta gritos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.72	0.72	0.72	802

Tabla 49 – Película "El Sexto Sentido" resultados etiqueta gritos

- Disparos:

Matriz de confusión:

		Valor Predicho	
		0	**1**
Valor Real	**0**	9	2
	1	2	9

Tabla 50 - Película "El Sexto Sentido" matriz de confusión etiqueta disparos

Los valores de precision, *recall* y *F-scores*:

	Precisión	**Recall**	***F-scores***	**Support**
Avg / total	0.82	0.82	0.82	22

Tabla 51 – Película "El Sexto Sentido" resultados etiqueta disparos

En esta película el clasificador ordena los valores obteniendo unos resultados que pueden considerarse válidos, ya que clasifica correctamente más del 80% de los valores.

4.11. Análisis de los resultados obtenidos.

Para visualizar mejor los resultados, los dividimos por las tres etiquetas que hemos utilizado y vemos en conjunto todas las películas, para ello calculamos la

media, varianza y desviación típica de los resultados. Estos valores los hemos obtenido a partir de la herramienta comentada en el apartado **Matlab**. Matlab utilizando las funciones *mean, var* y *std*, respectivamente.

- Gritos:

Película	Precisión	Recall	*F-scores*
Soy Leyenda	0.73	0.62	0.57
Salvar al Soldado Ryan	0.49	0.49	0.49
Piratas del Caribe y la Perla Negra	0.71	0.57	0.50
Independence Day	0.63	0.59	0.56
Fight Club	0.66	0.58	0.51
Armageddon	0.72	0.63	0.59
Eragon	0.65	0.54	0.42
Dead Poets Society	0.73	0.55	0.43
Billy Elliot	0.75	0.74	0.74
El Sexto Sentido	0.72	0.72	0.72

Tabla 52 – Resultados totales precisión, recall y *F-scores* para la etiqueta gritos.

Los resultados obtenidos para la etiqueta gritos son:

	Media	Varianza	Desviación Típica
Precisión	0.6790	0.0060	0.0774
Recall	0.6030	0.0061	0.0780

F-scores	0.5530	0.0118	0.1085

Tabla 53 – Resultados totales etiqueta gritos, valores de la media, varianza y desviación típica.

- Explosiones: [1]

Película	Precisión	Recall	*F-scores*
Soy Leyenda	0.88	0.88	0.88
Salvar al Soldado Ryan	0.63	0.60	0.58
Piratas del Caribe y la Perla Negra	0.85	0.85	0.85
Independence Day	0.87	0.87	0.87
Fight Club	0.85	0.80	0.79
Armageddon	0.82	0.73	0.70
Eragon	0.64	0.63	0.63
Dead Poets Society	-	-	-
Billy Elliot	-	-	-
El Sexto Sentido	-	-	-

Tabla 54 – Resultados totales precisión, recall y *F-scores* para la etiqueta explosiones.

[1] Los valores expresados en la tabla como "-" indican que dicha película no contiene el evento violento que estamos estudiando.

Los resultados obtenidos para la etiqueta explosiones son:

	Media	**Varianza**	**Desviación Típica**
Precisión	0.7914	0.0118	0.1085
Recall	0.7657	0.0132	0.1150
F-scores	0.7571	0.0147	0.1213

Tabla 55 – Resultados totales etiqueta explosiones, valores de la media, varianza y desviación típica.

- Disparos: [2]

Película	**Precisión**	**Recall**	***F-scores***
Soy Leyenda	0.80	0.79	0.79
Salvar al Soldado Ryan	0.64	0.54	0.68
Piratas del Caribe y la Perla Negra	0.67	0.55	0.46
Independence Day	0.88	0.87	0.87
Fight Club	0.71	0.67	0.65
Armageddon	0.87	0.85	0.85
Eragon	-	-	-
Dead Poets Society	-	-	-

[2] Los valores expresados en la tabla como "-" indican que dicha película no contiene el evento violento que estamos estudiando.

Billy Elliot	-	-	-
El Sexto Sentido	0.82	0.82	0.82

Tabla 56 – Resultados totales precisión, recall y *F-scores* para la etiqueta disparos.

Los resultados obtenidos para la etiqueta disparos son:

	Media	**Varianza**	**Desviación Típica**
Precisión	0.7700	0.0093	0.0966
Recall	0.7271	0.0196	0.1401
F-scores	0.7314	0.0212	0.1458

Tabla 57 – Resultados totales etiqueta disparos, valores de la media, varianza y desviación típica.

En este estudio, hemos observado, que obtenemos valores de *precisión*, *recall* y *F-scores* correctos como valores no tan válidos. Esto ha dependido de la película analizada, o incluso dentro de la misma película, dependiendo de la etiqueta observada obteníamos valores diferentes.

En el caso de guiarnos por los valores obtenidos en cada película, vemos que las películas "Soy Leyenda" y "Armageddon" son las que mayor valor de *precisión y recall* tienen. Esto puede deberse a que son películas con mucha acción, en las que se obtienen muchas muestras de las etiquetas estudiadas. También se pueden observar las gráficas *precisión-recall* en el ANEXO I – GRÁFICAS EXPERIMENTOS, donde se aprecian mejor estas variaciones.

Si nos guiamos por el tipo de etiqueta, la que peores resultados obtiene es la de gritos, es la que peor media obtiene para los tres conceptos que analizamos en la Tabla 53. La consecuencia de estos valores podría ser la difícil clasificación de

un grito, el cual puede durar muy poco, pero en las muestras que utilizamos se haya supuesto que dure más de lo debido.

Los valores obtenidos en este proyecto, en media, son valores razonables y coherentes. En relación con los estudios expuestos en el apartado 2.1, son unos valores peores que los que obtienen en esos estudios. Pero con los conocimientos y herramientas utilizados, los cuales eran los posibles, los valores pueden considerarse válidos para el proyecto.

En este estudio solo se ha utilizado un *kernel* lineal, porque era el más adecuado para trabajar con el tipo de valores que utilizamos. Pero habría otras posibilidades como el *kernel* gaussiano.

Otra posibilidad es trabajar con vectores multiclase, en este proyecto no se hizo por no alcanzar los resultados esperados al ejecutar el código del estudio.

5. CONCLUSIONES Y LÍNEAS FUTURAS

5.1. Conclusiones.

Hoy en día existen muchos estudios sobre la detección de ciertos sonidos en archivos de audio. Como presentamos anteriormente, hay algunos que intentan predecir ciertos sonidos que pueden considerarse como violentos, y obtienen valores muy altos en cuanto a la precisión en la predicción se refiere.

En este proyecto, como hemos visto en el apartado 4.11, obtenemos unos resultados totales a partir de los cuales podemos considerar que las herramientas utilizadas, así como los métodos de clasificación desarrollados en el proyecto son válidos para el estudio planteado, si bien deben mejorarse para poder optimizar en un futuro sus aplicaciones.

Los datos de los que partimos para iniciar el desarrollo del proyecto fueron obtenidos de la base de datos utilizada, donde se encontraban las películas estudiadas etiquetadas para los diferentes eventos violentos. Para un mejor desarrollo del proyecto se podría haber desarrollado un detector de audio, donde a partir del audio recogido se hubieran podido extraer las características de este y etiquetar el mismo.

Por otro lado, en este proyecto hemos desarrollado la clasificación para tres tipos de eventos violentos en el audio: gritos, explosiones y disparos. Pero además de estos eventos existen muchos otros que pueden caracterizar el audio, por ello sería interesante poder analizar más tipos de eventos violentos para así extender la clasificación utilizada en este proyecto.

En definitiva, el camino aquí estudiado para la clasificación de ciertos eventos violentos es válido, una vez establecidos los archivos de audio a tratar y extraídas las características de dichos audios, podemos establecer una clasificación correcta de estos eventos.

5.2. Líneas futuras.

En este proyecto hemos desarrollado una clasificación válida para ciertos eventos violentos, pero sería interesante continuar mejorando dicha clasificación y poder así obtener mejores resultados.

En primer lugar, se debería utilizar un mayor abanico de películas para el estudio, donde se encontrarán películas muy diferentes entre ellas. Una vez obtenidas, la extracción del audio y su posterior etiquetado en función de los eventos violentos que se estudien se pueden llevar a cabo como parte del proyecto, conociendo de esta manera las características propias de cada audio extraído.

En segundo lugar, podría realizarse el mismo estudio aquí expuesto, pero con diferentes métodos de clasificación supervisada, como por ejemplo k-NN *(Nearest Neighbours).*

Además de estas mejoras, habrá muchas otras en el futuro que busquen encontrar una mayor concordancia y precisión en lo que a los resultados se refiere.

6. BIBLIOGRAFÍA

[1] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio and M. Vento, *"Audio surveillance using a bag of aural words classifier," 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Krakow, 2013, pp. 81-86. doi: 10.1109/AVSS.2013.6636620.

[2] Enrique J. Carmona Suárez, (2014). *Tutorial sobre Máquinas de Vectores Soporte (SVM).* Universidad Nacional de Educación a Distancia (UNED) Madrid.

[3] Usman Malik, (2018). *Implementing SVM and Kernel SVM with Python's Scikit-Learn.* URL: https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/ Visitado por última vez: 20/09/2018.

[4] Mathworks, *"MATLAB, El lenguaje del cálculo técnico"*. URL: https://es.mathworks.com/help/matlab/ Visitado por última vez: 20/09/2018.

[5] Real Academia Española. *Diccionario de la Lengua Española.* URL: http://dle.rae.es Visitado por última vez: 20/09/2018.

[6] C. H. Demarty, C. Penet, G. Gravier, M. Soleymani, "*The MediaEval 2012 Affect Task : Violent Scenes Detection, in MediaEval 2012"* Workshop, ceur-ws.org, vol. 927, Pisa, October 2012

[7] C. H. Demarty, C. Penet, G. Gravier, M. Soleymani, *"A benchmarking campaign for detecting violent scenes in movies"*, ECCV2012 workshop on Information Fusion in Computer Vision for Concept Recognition, Firenze, October 2012.

[8] B. Ionescu, J. Schlüter, I. Mironică, M. Schedl, "*A Naive Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies*", ACM

International Conference on Multimedia Retrieval - ICMR 2013, Dallas, Texas, USA, April 16 - 19, 2013.

[9] Technicolor, URL: http://www.technicolor.com Visitado por última vez: 20/09/2018.

[10] C.H. Demarty,C. Penet, M. Soleymani, G. Gravier. *"VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation".* In Multimedia Tools and Applications, May 2014.

[11] C.H. Demarty, B. Ionescu, Y.G. Jiang, and C. Penet. *"Benchmarking Violent Scenes Detection in movies"*. In Proceedings of the 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), 2014.

[12] M. Sjöberg, B. Ionescu, Y.G. Jiang, V.L. Quang, M. Schedl and C.H. Demarty. "*The MediaEval 2014 Affect Task: Violent Scenes Detection*". In Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain (2014).

[13] C.H. Demarty,C. Penet, G. Gravier and M. Soleymani. "*A benchmarking campaign for the multimodal detection of violent scenes in movies"*. In Proceedings of the 12th international conference on Computer Vision – Volume Part III (ECCV'12),Andrea Fusiello, Vittorio Murino, and Rita Cucchiara (Eds), Col. Part III. Springer Verlag, Berlin.

[14] Scikit – learn. "*Machine Learning in Python".* URL: http://scikit-learn.org/stable/ Visitado por última vez: 22/09/2018.

[15] Scikit – learn. *"Support Vector Machine"*. URL: http://scikit-learn.org/stable/modules/svm.html .Visitado por última vez: 22/09/2018.

[16] MediaEval. "*MediaEval Benchmarking Initiative for Multimedia Evaluation".* URL: http://multimediaeval.org/ . Visitado por última vez: 23/09/2018.

[17] Davis, S.B., Mermelstein, P., *"Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences"*, IEEE Trans. on Acoustic, Speech and Signal Processing, 1980.

[18] Scikit-Learn. *"Choosing the right estimator"*. URL: http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html. Visitado por última vez 22/09/2018.

[19] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C.Martin, L.Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, *"The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data,"*in Affective Computing and Intelligent Interaction, 2007, vol. 4738, pp. 488–500.

[20] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *"DEAP: a database for emotion analysis using physiological signals"* IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[21] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, *"A multimodal database for affect recognition and implicit tagging,"* IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 42–55,Jan. 2012.

[22] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret and Liming Chen, *"LIRIS-ACCEDE: A Video Database for Affective Content Analysis".*

7. ANEXO I – GRÁFICAS EXPERIMENTOS

En este anexo se encuentran las gráficas de las curvas *precisión-recall* correspondientes a los experimentos realizados en el capítulo 4 del proyecto.

7.1. Película Soy Leyenda.

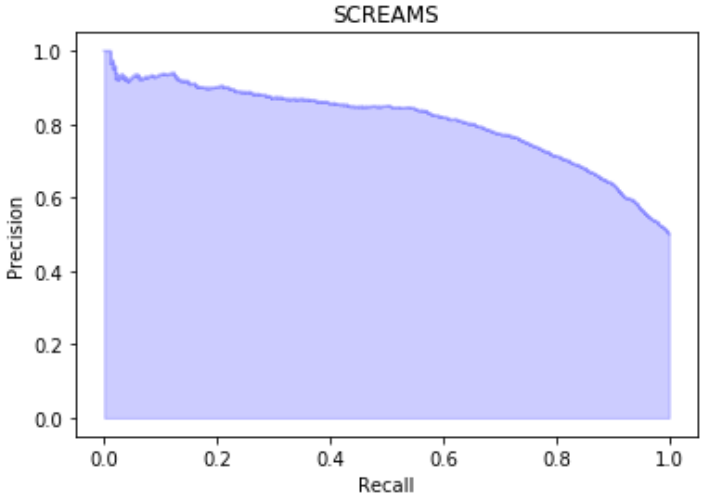

Figura 7 – Película "Soy Leyenda" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.8092

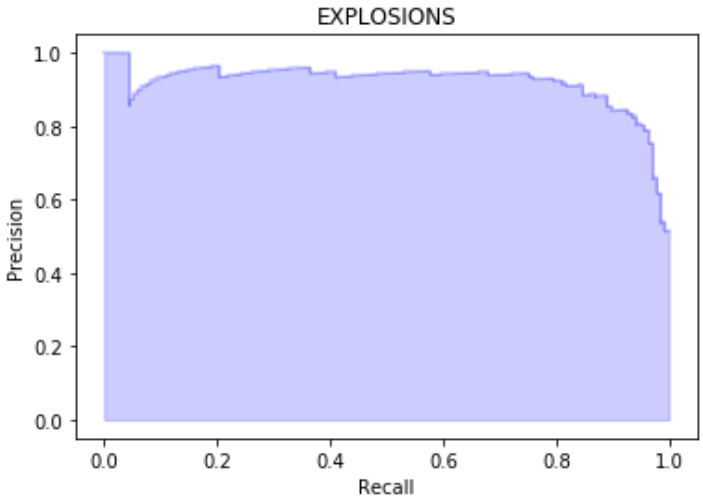

Figura 8 – Película "Soy Leyenda" curvas *Precision-Recall* etiqueta explosiones

Área bajo la curva = 0.9211

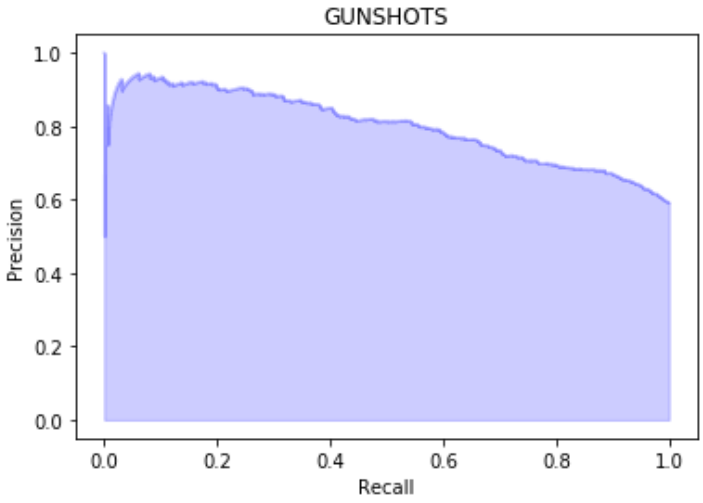

Figura 9 – Película "Soy Leyenda" curvas *Precision-Recall* etiqueta disparos

Área bajo la curva = 0.9055

7.2. Película Salvar al soldado Ryan.

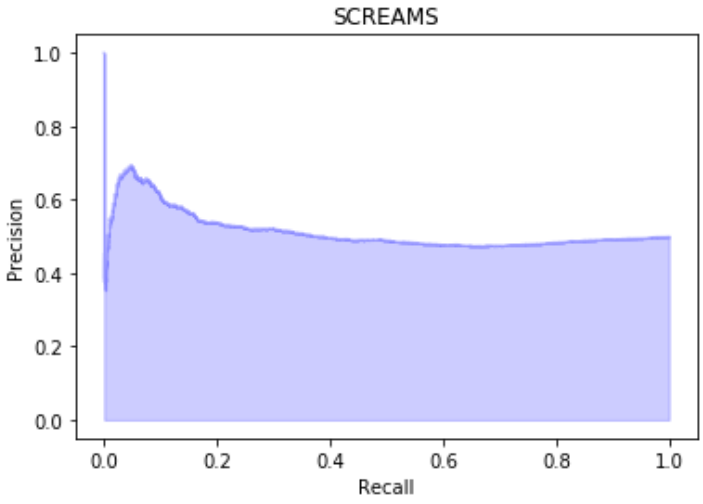

Figura 10 – Película "Salvar al soldado Ryan" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.5134

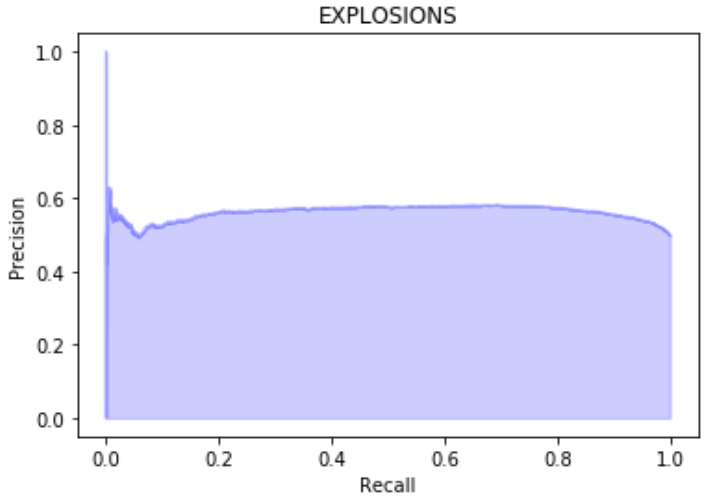

Figura 11 – Película "Salvar al soldado Ryan" curvas *Precision-Recall* etiqueta explosiones

Área bajo la curva = 0.5625

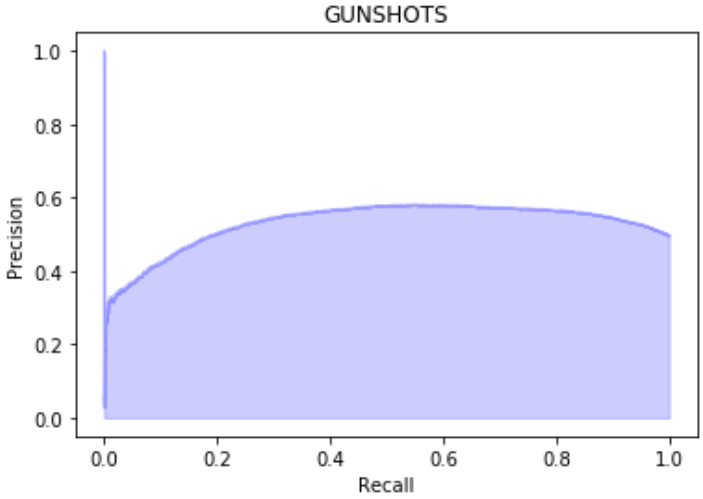

Figura 12 – Película "Salvar al soldado Ryan" curvas *Precision-Recall* etiqueta disparos

Área bajo la curva = 0.5296

7.3. Película Piratas del Caribe y la Perla Negra.

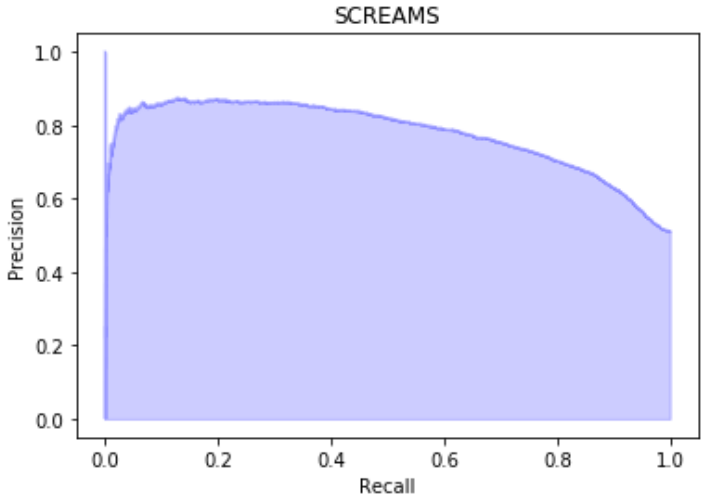

Figura 13 – Película "Piratas del Caribe y la Perla Negra" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.7777

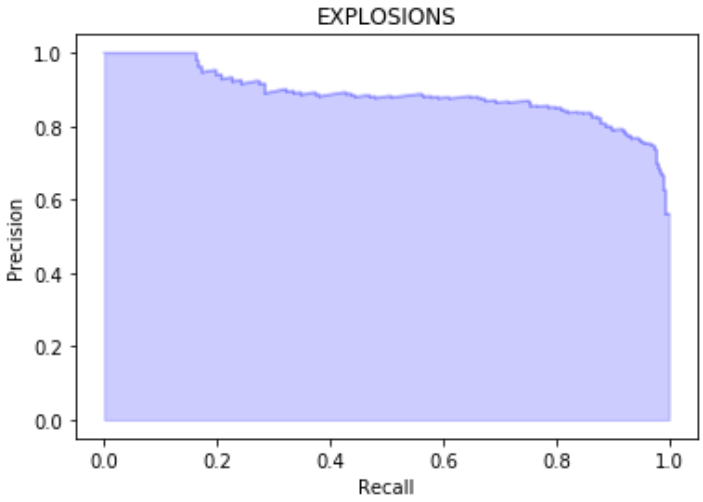

Figura 14 – Película "Piratas del Caribe y la Perla Negra" curvas *Precision-Recall* etiqueta explosiones

Área bajo la curva = 0.8870

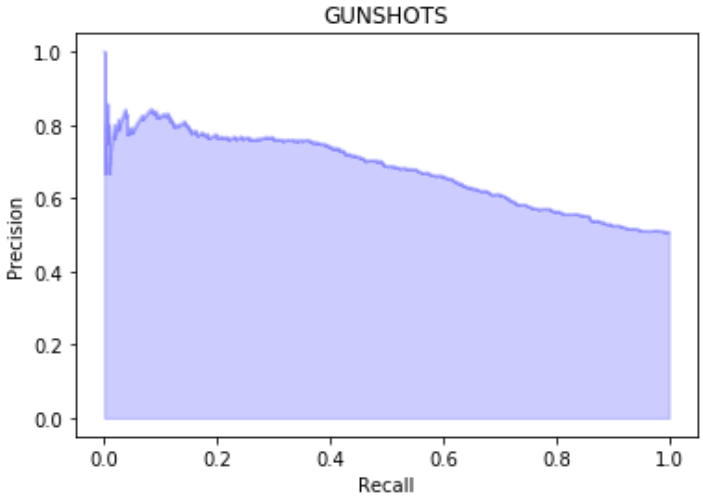

Figura 15 – Película "Piratas del Caribe y la Perla Negra" curvas *Precision-Recall* etiqueta disparos

Área bajo la curva = 0.6779

7.4. Película Independence Day.

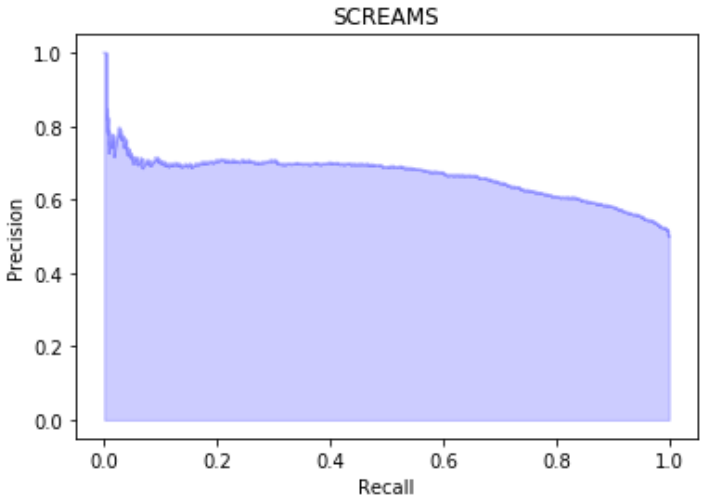

Figura 16 – Película "Independence Day" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.6654

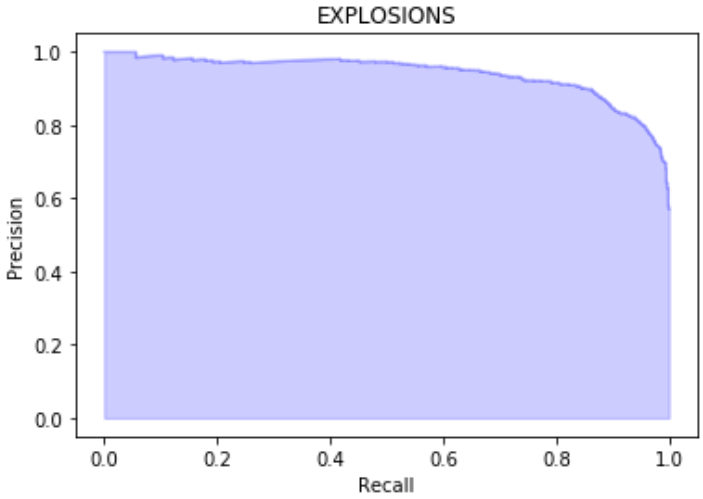

Figura 17 – Película "Independence Day" curvas *Precision-Recall* etiqueta explosiones

Área bajo la curva = 0.9421

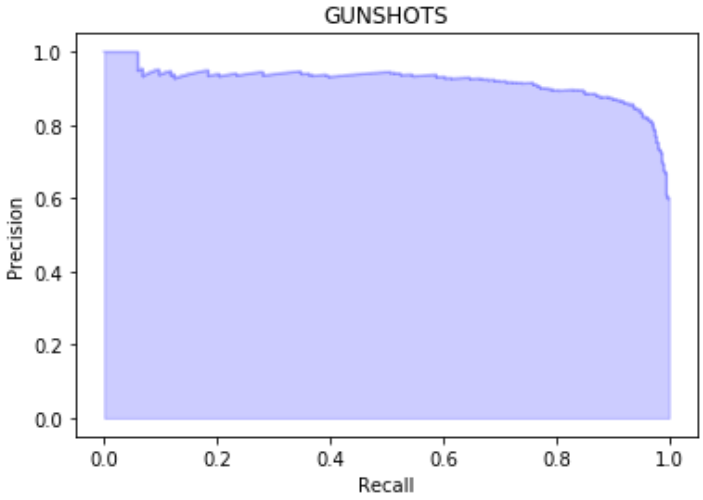

Figura 18 – Película "Independence Day" curvas *Precision-Recall* etiqueta disparos

Área bajo la curva = 0.9206

7.5. Película Fight Club.

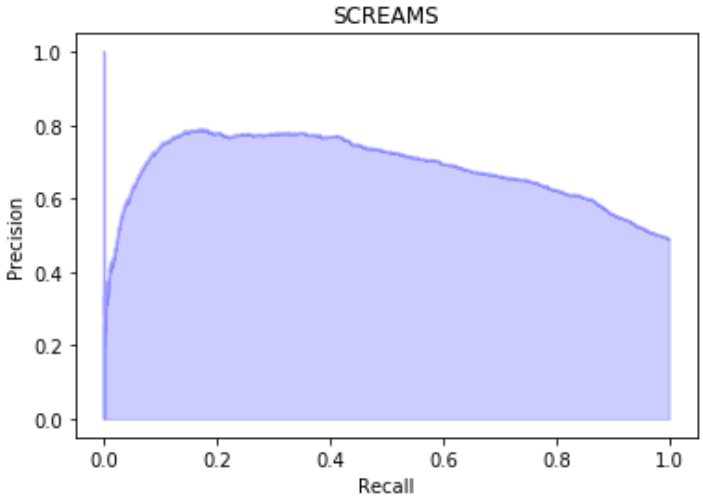

Figura 19 – Película "Fight Club" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.6812

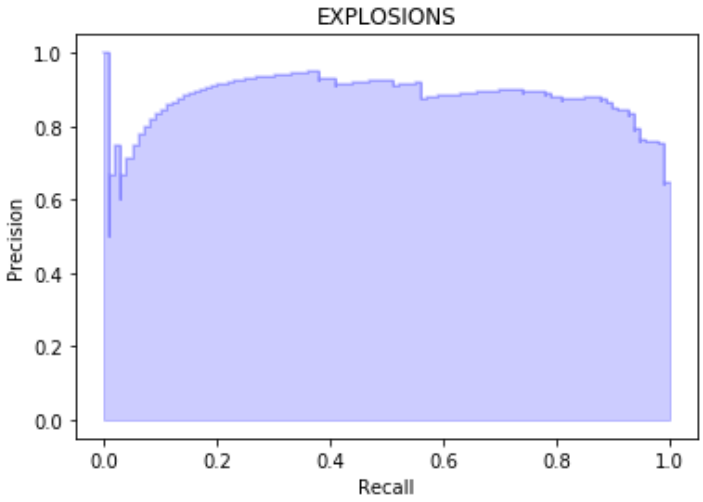

Figura 20 – Película "Fight Club" curvas *Precision-Recall* etiqueta explosiones

Área bajo la curva = 0.8762

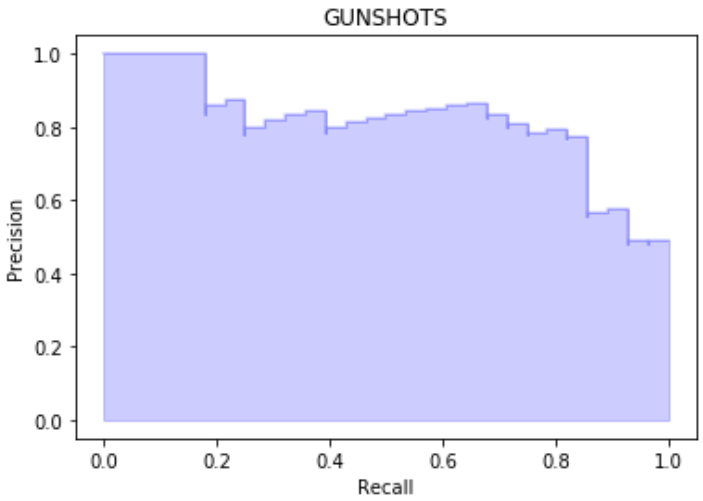

Figura 21 –– Película "Fight Club" curvas *Precision-Recall* etiqueta disparos

Área bajo la curva = 0.8107

7.6. Película Armageddon.

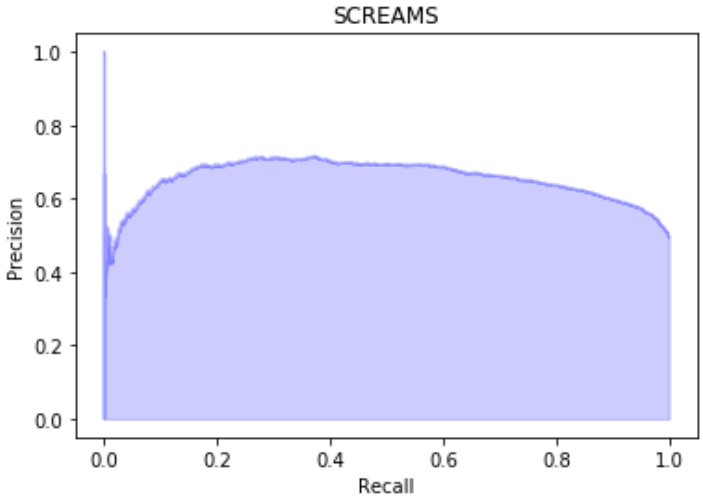

Figura 22 – Película "Armageddon" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.6530

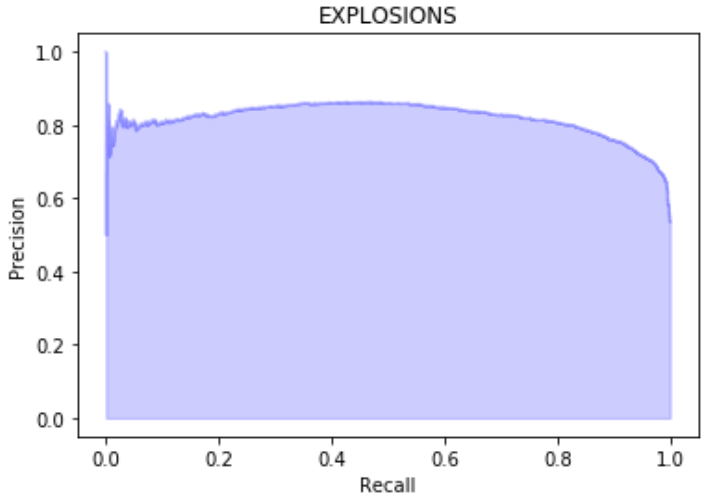

Figura 23 – Película "Armageddon" curvas *Precision-Recall* etiqueta explosiones

Área bajo la curva = 0.8187

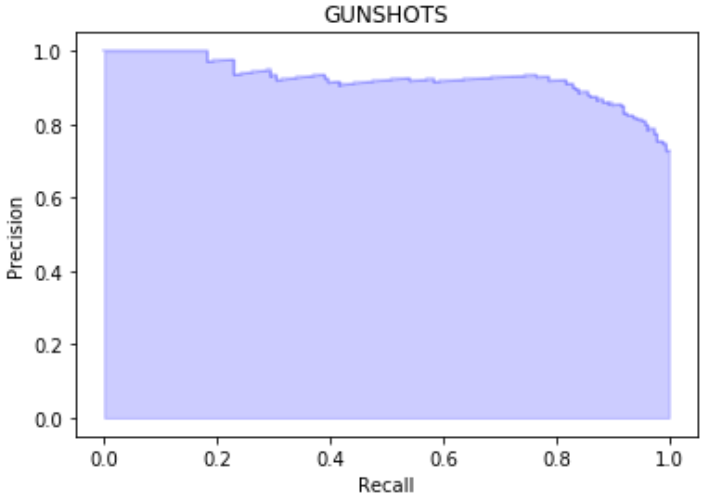

Figura 24 – Película "Armageddon" curvas *Precision-Recall* etiqueta disparos

Área bajo la curva = 0.9255

7.7. Película Eragon.

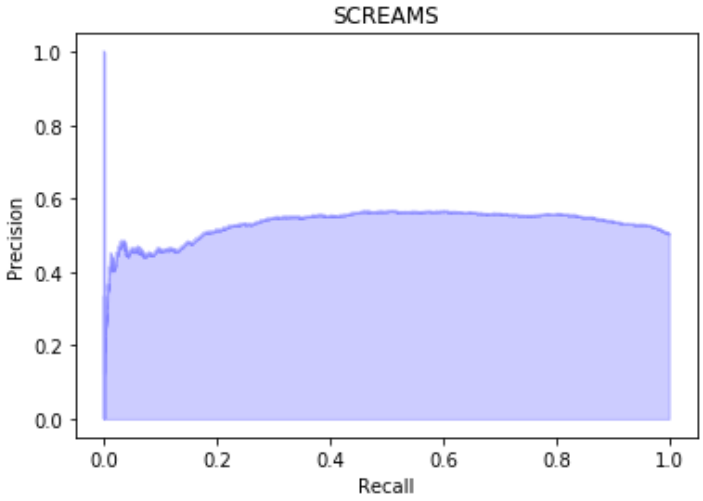

Figura 25 – Película "Eragon" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.5314

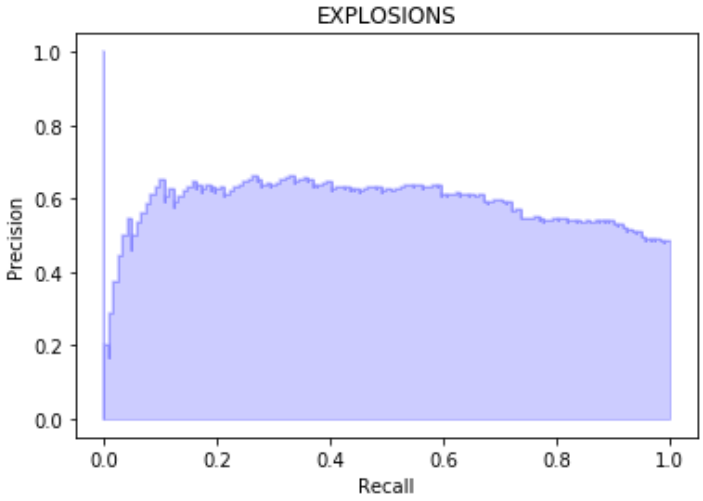

Figura 26 – Película "Eragon" curvas *Precision-Recall* etiqueta explosiones

Área bajo la curva = 0.5815

7.8. Película Dead Poets Society.

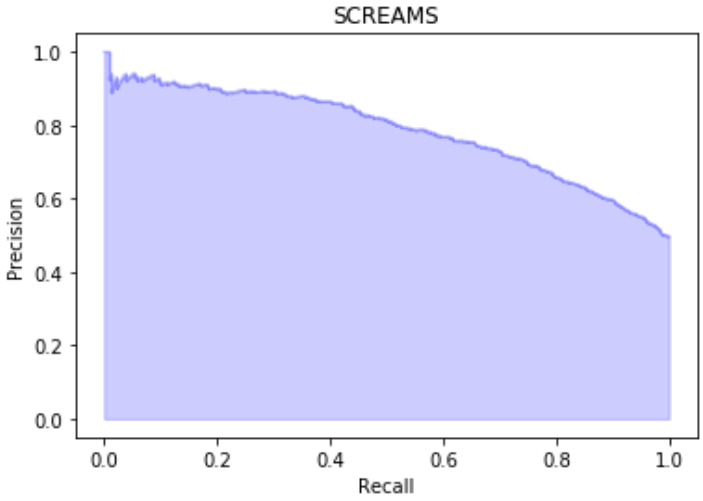

Figura 27 – Película "Dead Poets Society" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.7859

7.9. Película Billy Elliot.

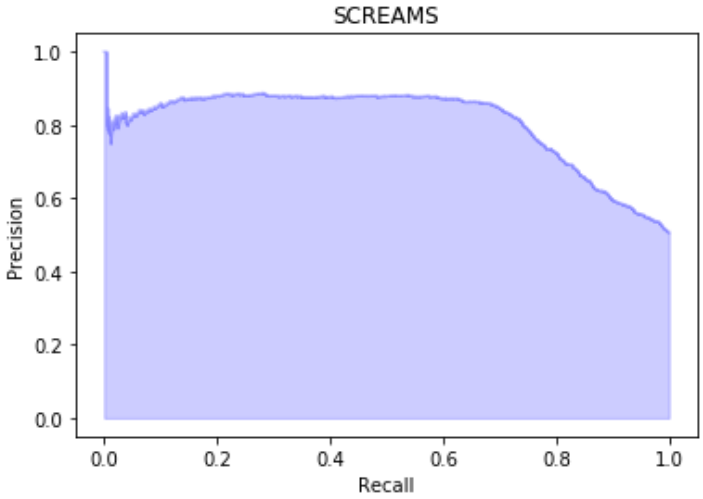

Figura 28 – Película "Billy Elliot" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.8085

7.10. Película El Sexto Sentido.

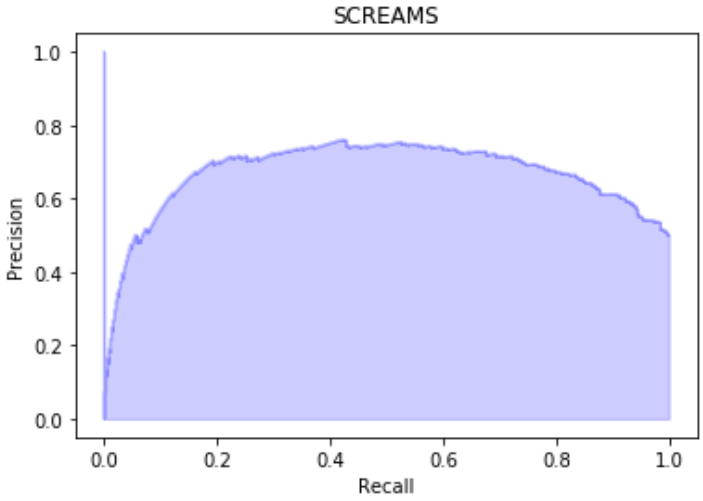

Figura 29 – Película "El Sexto Sentido" curvas *Precision-Recall* etiqueta gritos

Área bajo la curva = 0.6627

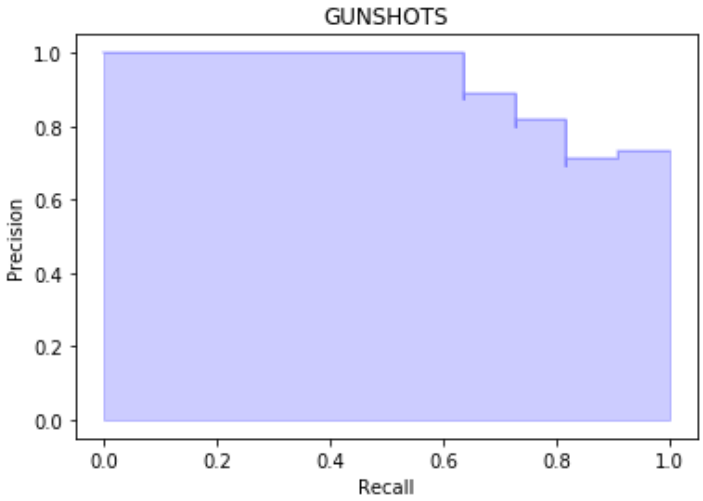

Figura 30 – Película "El Sexto Sentido" curvas *Precision-Recall* etiqueta disparos

Área bajo la curva = 0.9198

8. ANEXO II – PRESUPUESTO

En este apartado detallaremos la valoración económica del proyecto, donde expondremos el presupuesto. En primer lugar, dividiremos los costes en tres: coste de personal, coste del material y costes generales.

COSTES DE PERSONAL			
Cargo	**Horas Totales**	**Coste / Hora**	**Coste Total**
Profesor Titular – Ingeniero	60 horas	32,2 € / hora	1932 €
Estudiante	305 horas	8,2 € / hora	2501 €
TOTAL			**4433 €**

Tabla 58 – Costes de Personal

COSTES DE MATERIAL	
Tipo de Material	**Importe**
Ordenador Portátil	800 €
Licencia Matlab	69 €
Licencia Spyder	Licencia gratuita
TOTAL	**869 €**

Tabla 59 – Costes de Material

COSTES GENERALES	
Tipo	Importe
Electricidad	100 €
Conexión Internet	25 €
TOTAL	**125 €**

Tabla 60 – Costes Generales

Por tanto, el presupuesto total del proyecto es de 5427 €, que está desglosado en la Tabla 61.

PRESUPUESTO TOTAL	
Tipo de Coste	**Importe**
Costes de Personal	4433 €
Costes de Material	869 €
Costes Generales	125 €
TOTAL	**5427 €**

Tabla 61 – Presupuesto Total de Proyecto

9. ANEXO III - ABSTRACT.

9.1. Introduction and objetives.

Nowadays, movies, series, everything related to cinema are within everyone's reach. For that reason, it is interesting to know the effects or emotions that those movies or series will produce on us. Right now, all the movies are rated according to the age of the intended audience, It is assumed that if a movie is violent or has scenes that can produce strong emotions in people will be rated for older people. But the problem we are going to solve is: how are these films automatically rated? That is, how we can detect certain emotions in the audio of those films. Well, we are going to focus on discovering and trying to detect violent events, which are moments in movies that can produce fear, anxiety or anguish, for example: screaming, explosions, shooting; in said audio files.

Violence is one of the reasons for the arrangement of films according to age. The definition on this topic, which best fits this topic, is the one used by the World Health Organization (WHO); "Violence is the intentional use of force or physical power, in fact, or as a threat, against oneself, another person or a group or community, which causes or has a high probability of causing injury, death, psychological damage, development or deprivation." This definition of the concept speaks of the consequences of violence, which is closely related to the events we want to classify. These effects can produce feelings or emotions in people, normally violent events can produce fear or anxiety.

The main objective of this project is the detection of violent events in audio files. With this we want to be able to classify them and try to detect them from previously classified audio records, extracting their characteristics.

The audios used correspond to some well-known movies. From them a series of common and main characteristics of the audio have been extracted. Apart from these features we have some files in which the different events that can be extracted from these audios are classified.

9.2. Development of the proyect.

9.2.1. Theory.

In this section we will present the theoretical concepts used in the development of the project.

9.2.1.1. Machine Learning.

Machine learning is a type of artificial intelligence that develops in the machines or computers the ability to learn, without having been explicitly programmed. It focuses on the development of algorithms that give machines the ability to find certain patterns or behaviors, based on certain data or examples known a priori. The algorithms of machine learning are classified into two groups: supervised or unsupervised.

Supervised algorithms or supervised learning are those that, based on what has been learned about certain data previously labeled, are able to correctly label a data at the output of the system, that is, it is able to predict the output value.

Unsupervised algorithms or unsupervised learning take place when there is no previously tagged data, there is no a priori knowledge. The input data to the system are known, but there is no output data related to the input as it happened in the supervised learning. For this development you have to find some kind of organization or classification that helps simplify the analysis. A data structure must be found by observations.

We decided to use the Support Vector Machine or SVM algorithm because it has functions that adapt and help to obtain the results of the project. SVM is a set of supervised learning algorithms related to classification and regression problems. It consists of given a set of training samples are labeled the classes that exist in it and from them train an SVM that predicts the class of a new sample. The SVM looks for a hyperplane that optimally separates the points of one class or another. Here, in the concept of "optimal separation" is where the fundamental characteristic of the SVM resides, the

hyperplane sought will be the one that provides the maximum separation the points that are closest to it.

Once the SVM algorithm is defined, we use the precision - recall curves to visualize the classification of said algorithm. These curves are used to evaluate the output quality of the classifier. In information retrieval, accuracy is a measure of the relevance of the result, while recall is a measure of how many results are returned. To understand it better, we must define previously some concepts:

- TN or True Negatives: The number of times the outcome of a negative event is classified as negative.
- TP or True Positive: The number of times the outcome of a positive event is classified as positive.
- FN or False Negatives: The number of times the outcome of a positive event is classified as negative.
- FP or False Positives: The number of times the outcome of a negative event is classified as positive.

Therefore, once these definitions are seen, we will clarify the concepts of precision and recall.

The precision for a class is the number of true positives, that is, the number of elements that have been correctly labeled as belonging to the positive class, divided by the total number of elements labeled as belonging to the positive class.

Recall is defined as the number of true positives divided by the total number of elements that belong to the positive class, that is, the sum of true positives and false negatives, which are elements that were not labeled as belonging to the positive class, but they should be.

Another parameter that is used to evaluate the accuracy of a classification model is *F-scores*. It is defined as the geometrical average of the precision and recall. *F-scores* takes values between zero and one, with a value close to or equal to one corresponding to a correct classification, where accuracy and

recall is almost perfect. If it takes a value close to or equal to zero, the model is not correct.

9.2.2. Implementation.

In this section we will detail the process of obtaining the results of the project.

9.2.2.1. Databases.

At the beginning of the project different databases were evaluated that would be of help for the extraction of certain data that would allow the study that we have developed.

▪ TECHNICOLOR database: composed of 25 movies and 86 YouTube videos. It includes a file that information about when violent events occur in the movie, as well as two files the characteristics of each movie.

▪ MAHNOB-HCI database: composed of 20 excerpts from films, whose duration ranges from 35 to 118 seconds. It contains labels related to emotions, excitement and predictability combined with videos of people's facial expressions. These data were collected in a total of 30 participants while viewing the movies and videos provided by the database.

▪ DEAP database: contains 120 videos of approximately one-minute duration. It is a multimodal database for the analysis of the moods of human beings. Through an electroencephalogram and the physiological signals of 32 participants, they characterize the emotions produced in people and their level.

▪ HUMAINE database: composed of 50 videos with a duration ranging from 5 seconds to 3 minutes. It identifies emotional states, key events, words related to emotions.

▪ LIRIS-ACCEDE database: composed of 9800 excerpts from 160 films, lasting between eight and twelve seconds. Excitation and valence measurements are labeled.

9.2.2.2. Modeling the data.

We decided to carry out the project using the TECHNICOLOR database, since it is the one that best fits the data we need, it contains several files that help us develop the project.

In this database we have several .txt files in which certain events that have occurred throughout the film are tagged. We have a total of eleven files, but not all of them are of interest. Therefore, we decided to make another small selection and consider only the the files that best fit the data we need. Therefore, we use the files related to violent events, which are shouting, shooting and explosions.

Next, we make a transformation to the data that we have extracted from the database. For this we perform operations in the Matlab tool. The data obtained from the database are imbalanced, there are many more samples of one class than another. To solve this problem we use the scikit-learn functionalities to correct the imbalanced data, for it there are two main techniques: Over Sampling and Under Sampling Over Sampling is to increase the number of samples of the class with the lowest number of samples until both classes have the same number. Under Sampling, however, is based on reducing the number of samples in the class with the largest amount, until it has the same number of samples as the other class. The technique we have used in this project is that of Under Sampling.

Once we have the data correctly modeled, we continue with the implementation of the algorithm chosen to perform the classification, the Support Vector Machine or SVM. We will also represent the precision-recall curves and calculate the area under said curve.

9.2.3. Experimentation and results.

In this section we will expose the results obtained when performing the experiments, for this we will rely on the concepts precision, recall and *F-scores*.

To visualize better the results, we divide them by the three labels that we have used, and we see all the films together, for this we calculate the mean, variance and standard deviation of the results. We have obtained these values from the tool commented in the Matlab section. Matlab using the mean, var and std functions, respectively.

- Screams:

	Mean	**Variance**	**Standard Deviation**
Precision	0.6790	0.0060	0.0774
Recall	0.6030	0.0061	0.0780
F-scores	0.5530	0.0118	0.1085

Tabla 62 – Results Label Screams

- Explosions:

	Mean	**Variance**	**Standard Deviation**
Precisión	0.7914	0.0118	0.1085
Recall	0.7657	0.0132	0.1150
F-scores	0.7571	0.0147	0.1213

Tabla 63 – Results Label Explosions

- Gunshots.

	Mean	**Variance**	**Standard Deviation**
Precisión	0.7700	0.0093	0.0966

Recall	0.7271	0.0196	0.1401
F-scores	0.7314	0.0212	0.1458

Tabla 64 – Results Label Gunshots

9.3. Conclusion.

Today there are many studies on the detection of certain sounds in audio files. As we presented previously, there are some that try to predict certain sounds that can be considered as violent and obtain very high values in terms of prediction accuracy.

In this project, as we have seen in section 4.11, we obtain total results from which we can consider that the tools used, as well as the classification methods developed in the project, are valid for the proposed study, although they should be improved for to be able to optimize their applications in the future.